

THIS WEEK

EDITORIALS

EDITORIAL The human genome sequence at ten. What happens now? **p.140**

WORLD VIEW Lab closure shows we must stand up to drug firms **p.141**

HE'S THE ONE Female finches stressed by poor mates **p.142**



Letting the bugs out of the bag

The public should be properly consulted ahead of any release of experimental insects. But what do they need to know, and whose job is it to ensure the message gets across?

In the week before Christmas, some 6,000 genetically modified (GM) mosquitoes were deliberately introduced to an uninhabited forest in Malaysia. The move took many local people and international observers by surprise. They had thought that the trial, which aims to investigate how long the modified insects live and how far they can fly, had been postponed.

The mix-up was down to the media confusing the trial with a second planned experiment, due to take place in a populated area later this year. But it adds to a growing sense of unease among some in the field about the way in which the public are consulted and notified about such experiments. The Malaysian trial, developed as an approach to controlling dengue fever by the British biotech company Oxitec, based in Oxford, followed the release of 3.3 million of the firm's GM insects in separate tests in the Cayman Islands in 2009 and 2010.

There is no suggestion that any of the releases was unsafe, or contravened any law. In line with Malaysia's biosafety rules and the Cayman Islands' draft rules, permits were issued after the relevant national authorities performed risk assessments.

But scientists and local people alike have taken issue with the manner in which the public engagement was handled, as well as the choice of the Cayman Islands, where, unlike Malaysia, biosafety considerations are not well developed. Even specialist researchers in the GM mosquito field — hardly a sprawling sector — say that they first heard about Oxitec's experiments in the Cayman Islands only when the company announced the results at an academic conference in November.

If the release of GM organisms is handled badly, it could generate an unnecessary and unhelpful climate of suspicion. One problem is that there is no standard laboratory procedure when it comes to informing the public of such experiments. Moreover, is merely informing them sufficient? Given the farce over the use of GM crops in Europe, early buy-in and support from local communities would be a good way to deflect unfounded fears that could surface in the future, particularly given that early findings are promising. (Oxitec says the release of the GM mosquitoes in the Cayman Islands study successfully reduced the wild dengue-carrying population by about 80%.) But researchers who work on GM insects say that they are unsure how much public engagement is enough and who has responsibility for it.

Transparency is essential. The Malaysian authorities went to some lengths to inform people that the trials were going ahead, holding open forums and briefing the media, which gave the experiments wide coverage. The resulting discussion highlighted concerns. It also seeded an appetite for more information, which seems to have been responsible for the subsequent confusion over the trial's timing. By contrast, efforts by the Cayman Island authorities seem to have amounted to not much more than producing little-reported leaflets and a video, posted on YouTube and broadcast on television, which failed to say that the mosquitoes were genetically modified — the main concern of critics.

Researchers, both in the public and private sector, should do more to ensure that the relevant authorities make the relevant facts available, or do so themselves. It is they, not the authorities, after all, that will probably be the focus of protests and complaints if public engagement is handled badly. With this in mind, scientists at the University of California, Irvine,

"If the release of GM organisms is handled badly, it could generate an unnecessary and unhelpful climate of suspicion."

have developed and published a detailed and ambitious framework to engage the public in global-health initiatives (J. V. Lavery *et al. Trends Parasitol.* **26**, 279–283; 2010) — heavily based on their own experiences with GM mosquito research in Mexico.

In the absence of guidelines to help researchers to deal with local communities, the authors produced 12 of their own, which include rigorous site selection, to ensure that the purpose and goals of the research are made clear, and the use of focus groups and citizen councils to probe local opinions and to decide whether informed consent is necessary. Although many of the issues are common to such research, the decisions must be taken on a site-by-site basis, they say. The World Health Organization is also drawing up guidelines, which it says will help scientists to assess the social and cultural issues relating to their work.

Oxitec acknowledges that there are lessons to learn from its experiences. Best placed to judge the results of this are the people of Brazil, the planned site of the company's next experiment.

So far, GM mosquitoes and other insects have largely flown beneath the radar. That will change sooner or later. It is surely better that the scientists involved bring them to the public's attention, rather than have that attention thrust upon them by others. ■

A fair share

The Hungarian government needs to up its stake in the nation's scientific future.

For almost 20 years, the Collegium Budapest has stood as a symbol of a new era of science in central and eastern Europe. Some 700 scholars from 40 countries have spent time in its rarefied intellectual atmosphere — an esteemed institute for advanced study — where, free from teaching and administrative burdens, they have produced hundreds of papers and books in fields ranging from economics to political sciences, theoretical biology and the humanities. Given its widely recognized success, why does the collegium now face threats to its survival?

On one level, its problems are financial. The institute's international sponsors, including a number of western European governments, banks and private foundations, want the Hungarian government to bear much more of the collegium's annual cost, which runs to around €1.2 million (US\$1.6 million). The Hungarian hosts currently contribute just €100,000 per year, which goes towards the costs of accommodation and salaries for a core staff of 30 or so visiting and permanent fellows. This is little more than it paid during the relatively lean early years after the institute opened in 1992. Given the country's expanded economic potential and its membership of the European Union (EU) since 2004 — of which it currently holds the presidency — the German government and other sponsors have asked Hungary to boost its share to about half the annual costs. If Hungary does not find the money, foreign sponsors say that they will withdraw their support.

But €600,000 is apparently more than Hungary is willing to pay for an intellectual enclave of international reputation, housed in the former city hall in Budapest's historic castle district, provided rent-free by the Hungarian Academy of Sciences.

Miklós Réthelyi, Hungary's minister for national resources (and science), promised last year to examine whether EU structural funds could be used to maintain the collegium. But in December, discussion of the issue was again postponed, adding to concerns that Hungary is no longer interested in keeping the institute alive. The collegium's assembly of members, which will discuss the collegium's future at a meeting in April, is beginning to lose hope.

Perhaps the Hungarian government would not be particularly sorry to lose this academic jewel. The collegium is also known as a haven of outspokenness, and some suspect that the output of some of its

scholars is unwelcome in government circles. Hungarian economist János Kornai, for example, recently published a caustic analysis of current political tendencies in Hungary.

If the collegium is forced to close, much will be lost. Institutes for advanced study are a vital element of modern science systems — a niche in the bustle of academic routine where researchers can find the time to elaborate on thoughts and concepts, and exchange ideas with colleagues from other disciplines. In Hungary, the Collegium

“The political and societal challenges ahead certainly demand honesty and trust.”

Budapest brings an international flavour to Hungarian science. Senior figures from overseas are hard to find in its other universities and research institutes.

The Hungarian government should have the courage to do the right thing and take on a fair share of the costs, even if it doesn't primarily serve current domestic needs. Doing so would help to counter the widespread impression — furthered by a new and restrictive media law, and by a badly handled row over alleged misuse of research grants by a group of philosophers with the Academy of Sciences — that Hungary's leadership is drifting towards autocracy and that critical discourse is being stifled. The political and societal challenges ahead certainly demand 'honesty and trust' — the title of a Collegium Budapest project on the post-socialist transformation process. Budapest, with its rich scholarly tradition, has been an ideal place for people to study and reconcile diverging cultures of knowledge in a reshaped Europe. The changes under way in the Arab world may reshape East–West relations on a much larger scale. The Collegium Budapest would be a good place to begin to ponder what that might mean. ■

Best is yet to come

Ten years after the human genome was sequenced, its promise is still to be fulfilled.

Former US president Bill Clinton called it the “most important, most wondrous map ever produced by humankind”. To then UK prime minister Tony Blair, it was a “breakthrough that takes humankind across a frontier and into a new era”. His science minister David Sainsbury said: “We now have the possibility of achieving all we ever hoped for from medicine.” When *Nature* published a 62-page article on 15 February 2001 titled ‘Initial sequencing and analysis of the human genome’ it is not difficult to see why the world got excited. Perhaps, even, a little overexcited. One of our editors, Henry Gee, penned a newspaper piece at the time that promised, by 2099, “genomics will allow us to alter entire organisms out of all recognition, to suit our needs and tastes... [and] will allow us to fashion the human form into any conceivable shape. We will have extra limbs, if we want them — maybe even wings to fly.”

As Eric Lander, director of the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, and the first author on that 2001 paper, writes on page 187 of this issue: “The human genome has had a certain tendency to incite passion and excess.” A decade on, Lander notes, the pattern continues, with “a front-page news story on the tenth anniversary of the announcement that chided genome scientists for not yet having cured most diseases”. The 2001 sequence was always a milestone on the journey to better medical care, rather than a destination. The ten-year anniversary of the publication in *Nature* and *Science* of sequences prepared respectively by the international Human Genome Project and Celera Genomics, now of Alameda, California, provides another — as well as an opportunity to reflect on progress.

Some things have undoubtedly changed. *Nature's* Editorial page in the 15 February 2001 issue examined not the scientific and medical

promise of the genome sequence, but the challenge of public access to information gathered by the commercial genomics sector. Acrimony over the differing public and private approaches has since faded; concerns over access to genomic data now centre on privacy issues.

Has medical progress been slower than was expected at the time? In an article on page 204, Eric Green and Mark Guyer of the US National Human Genome Research Institute in Bethesda, Maryland, offer an “updated vision” of the prospects for genomic medicine. “Significant change rarely comes quickly,” they write. “Although genomics has already begun to improve diagnostics and treatments in a few circumstances, profound improvements in the effectiveness of healthcare cannot realistically be expected for many years.” Research is not enough, they say, and new policies and practices as part of an expanded global effort are needed too.

The sequencing of the human genome was in many ways a triumph for technology as much as it was for science. That technology has continued to develop over the past decade, which Elaine Mardis of the Genome Center at Washington University in St Louis describes in an article starting on page 198 as a “remarkable sequencing technology explosion”.

Massively parallel sequencing technology allows questions to be asked and answered with “unprecedented speed and resolution”, she says. “The continuing upward trajectory of sequencing technology development is enabling clinical applications that are aimed at improving medical diagnosis and treatment.” A useful example is the development of genome-wide association studies to probe the underlying genetic landscape of some common diseases.

More than a decade ago, Michael Dexter, then head of the UK Wellcome Trust, which took part in the Human Genome Project, branded the genome sequence as the outstanding achievement of human history, eclipsing the significance both of the Moon landings and of the invention of the wheel. It is too early for that history to be written. For the genome sequence to be a true success, we must yet ensure that greater achievements are built on it. ■

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunq



Pharmaceutical industry must take its medicine

To fix the drug pipeline, governments must take on drug-makers instead of capitulating to their every demand, says Colin Macilwain.

Pfizer's announcement last week that it is to pull the plug on its drug-development laboratory in Sandwich, Kent, and fire most of its 2,400 staff (see page 154), must be a wake-up call for scientists and policy-makers alike. The pharmaceutical industry is taking them for a ride. Drug executives know that, however they behave, public money will continue to flow into the industry from spending on basic research and the purchase of final products.

For almost a decade now, drug-makers such as Pfizer have claimed that they can maintain huge research and development expenditures despite the increasing rarity of new 'blockbuster' drugs. This serves two purposes: it has persuaded investors that there is, really, something lucrative in the pipeline; and it has beguiled politicians into throwing public money at the early stages of drug development.

The closure of the labs in Sandwich is a sure sign that this process isn't delivering, in Britain or elsewhere. That is despite massive government investment — notably from the US National Institutes of Health, whose US\$32-billion budget is chiefly devoted to finding ideas for the industry.

Big pharma's fashionable younger brother, biotechnology, is not doing much better. It is experiencing the deepest and most prolonged slump in its 35-year history. When the most successful US biotechnology company, Amgen of Thousand Oaks, California, is taken out of the picture, the industry has never made a profit, as Gary Pisano, who studies technology strategy at Harvard Business School in Boston, Massachusetts, showed in his book *Science Business* (Harvard Business School Press, 2006). The 2010 report *How to Compete and Grow: A Sector Guide To Policy*, released by the McKinsey Global Institute in New York, found that biotechnology is unlikely to generate significant job growth. And *The Bioeconomy to 2030*, published by the Organisation for Economic Co-Operation and Development in Paris in 2009, noted that 75% of the economic impact of the life sciences is likely to be outside the health sector.

Yet the main thrust of scientific and regulatory policy in both Europe and the United States for ten years or more has been to give the leaders of the 'life-sciences industry' whatever they want, in the expectation that they will generate export earnings and highly paid jobs.

The most visible current features of British and US biomedical research policy are a pair of publicly funded megaprojects aiming to remove blockages in the drug pipeline. The planned UK Centre for Medical Research and Innovation in London and the proposed National Center for Advancing Translational Sciences at the US National Institutes of Health in Bethesda, Maryland, have their merits; those of the latter project were

spelled out by Gareth Fitzgerald in this space, last December.

But the political architects of these projects are applying their attention to the wrong part of the plumbing. It isn't just the stretch of pipeline that translates laboratory findings into drug candidates that is failing; it is drug development itself. If we want better value from investment in health research — not to mention the immense expenditure on drug treatments — then we need to upend the drug industry's operating model.

Policy-makers should look again at control of intellectual property and regulation. The grip of patenting on the life sciences has tightened, particularly since the World Trade Organization's international Trade-Related Aspects of Intellectual Property Rights agreement came into full force a decade ago. This tightening is what the industry wanted — it has bolstered profits and reduced drug piracy — but there is little evidence that it has increased the flow of innovative therapies.

More free exchange of information would be awkward, and innovation models such as that of the computer industry, where patented ideas are constantly swapped and resold, cannot be directly applied to drug development. However, many scientists — including, one suspects, the Pfizer staff too scared to talk to the BBC in Sandwich last week — are fed up with the secrecy and inefficiencies of the existing system, best exemplified by the fact that many clinical trials data never see the light of day. The regulatory system, meanwhile, is often blamed by the pharmaceutical industry for its problems — but actually serves the industry well, by setting up high barriers to entry.

Alternative approaches have been suggested. *The Manchester Manifesto*, published in November 2009 by a group led by John Sulston, a biologist, and Joseph Stiglitz,

an economist, both at the University of Manchester, UK, called for a new approach to the sharing of knowledge and data. Joyce Tait, a policy analyst at the University of Edinburgh, UK, has argued that a more flexible regulatory system (enabling, for example, drug trials on patient subgroups selected for their genetic susceptibility to certain treatments) could open the field to more players.

Scientists haven't embraced such possibilities aggressively enough, and politicians have barely engaged with them at all. They prefer to look to industry for advice on research and regulatory policy, and then beg it for favours. UK Prime Minister David Cameron even said in a speech last month that he had called Ian Read, Pfizer's chief executive, to inform him of yet another planned tax break, exempting revenue earned from patents held in Britain from corporation tax. His reward? Another 2,000 people unemployed. ■

Colin Macilwain is a contributing correspondent with Nature.
e-mail: cfmworldview@gmail.com

THE MAIN THRUST OF
POLICY
HAS BEEN TO GIVE
THE INDUSTRY
**WHATEVER IT
WANTS.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/jolkwh

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

GEOSCIENCE

Hard rains and stormy winds

North Atlantic hurricanes and their atmospheric remnants are the dominant cause of extremely heavy rainfall across vast swathes of the United States — as far north as Maine, and as far inland as Illinois.

Mathew Barlow of the University of Massachusetts at Lowell compared a data set of storm tracks and size with daily observations made between 1975 and 1999 at almost 9,500 weather stations in Central and North America. Over large areas of the northeastern United States, more than two-thirds of extreme precipitation events — rainfall exceeding 100 millimetres per day — were meteorologically related to hurricane activity occurring as far away as 500 kilometres. The strength and range of the storms' effects varied according to factors such as maximum wind speed.

Geophys. Res. Lett. doi:10.1029/2010GL046258 (2011)

ZOOLOGY

Mate mismatch causes stress

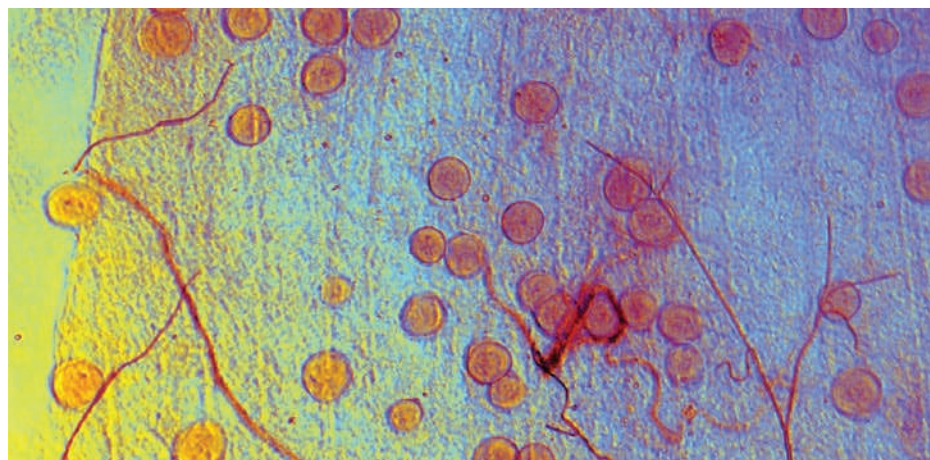
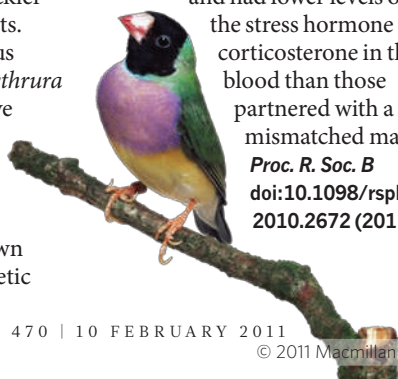


Female Gouldian finches that fail to land their ideal mate seem to have higher levels of stress than their luckier counterparts.

The monogamous Australian finches (*Erythrura gouldiae*, pictured) have either black or red heads, and females prefer to mate with partners whose head colour matches their own — an indication of genetic

compatibility. Simon Griffith of Macquarie University in Sydney, Australia, and his colleagues monitored the birds as they either chose their mates or were placed in a mating pair.

In both conditions, females that ended up with compatible males laid their first egg earlier and had lower levels of the stress hormone corticosterone in their blood than those partnered with a mismatched mate. *Proc. R. Soc. B* doi:10.1098/rspb.2010.2672 (2011)



PUBLIC HEALTH

Malaria mosquito lurks outdoors

The discovery of a new subgroup of malaria-carrying mosquito may explain why malaria eradication efforts have had limited success.

Malaria is caused by *Plasmodium* parasites (pictured, in red), which are transmitted by mosquitoes, mainly of the species *Anopheles gambiae*. Kenneth Vernick at the Pasteur Institute in Paris and his colleagues discovered the new subgroup of *A. gambiae* — dubbed Goundry for the village in the African country of Burkina Faso, where it was found — after collecting mosquito larvae from puddles,

raising them to adulthood in the lab and genetically analysing them.

Not only do the Goundry mosquitoes live primarily outdoors, where they avoid indoor insecticide sprays, they also acquire the parasite more easily than their indoor relatives. When fed with malaria-infected blood, 58% of Goundry mosquitoes picked up the disease, compared with just 35% of indoor mosquitoes.

Science 331, 596–598 (2011)

For a longer story on this research, see go.nature.com/7kzztq

NANOSCIENCE

Glimpses of crystal growth

One downside of electron microscopy is that the electron beam can damage the materials being imaged by breaking bonds and changing molecular structures. But Jamie Warner and his colleagues at the University of Oxford, UK, used this to their advantage, and obtained unprecedented pictures of crystals forming at the atomic level.

They directed an 80-kilovolt electron beam through a thin film of 'peapods' — carbon nanotubes containing spheres

of carbon atoms called buckyballs. Each buckyball contained two atoms of praseodymium. Prolonged exposure to the beam caused the buckyballs to coalesce, forming an inner nanotube. The praseodymium atoms were released into this inner nanotube, allowing them to form praseodymium carbide crystals.

The images reveal that the crystals formed as a result of atoms coalescing into clusters, which, in turn, clumped together, rather than by atoms being added one at a time to a single growing crystal. *ACS Nano* doi:10.1021/nn1031802 (2011)

TISSUE ENGINEERING

Blood vessels on demand

Growing blood vessels for use in cardiovascular surgeries is a tricky business. Vessels can be created from a patient's own cells, but the process is costly and takes up to nine months. Now researchers have devised a method that can churn out tens of vessels per donor that could then be stored until needed.

Shannon Dahl at Humacyte in Durham, North Carolina, and her collaborators grew their vessels by introducing the cells into scaffolds made of polyglycolic acid. Once these vascular grafts had grown in a bioreactor, the team stripped them of cells, reducing the likelihood of the vessels eliciting an immune response in the recipient.

The resulting collagen-tube grafts had similar properties to normal human blood vessels. When tested in a small number of baboons and dogs, most of the grafts remained open over test periods ranging between one month and one year.

Sci. Transl. Med. 3, 68ra9 (2011)
For a longer story on this research, see go.nature.com/9gyzoz

PHYSIOLOGY

Nitrate ups cells' efficiency

Leafy vegetables are chock-full of nitrate — a molecule that seems to boost the efficiency of energy-producing organelles called mitochondria in muscle cells.

Filip Larsen, Eddie Weitzberg and their group at the Karolinska Institute in Stockholm gave 14 volunteers either doses of nitrate similar to those found in certain foods or a placebo. Mitochondria collected from the muscle cells of those who had been taking nitrate for three days made 19% more energy-dense ATP molecules per oxygen molecule consumed than did those on the placebo.

Mitochondria must

maintain an electrochemical gradient across their inner membrane to produce ATP. Levels of a protein that saps these organelles' conductivity were reduced in people taking nitrate, suggesting that their muscle cells were producing energy more economically.

Cell Metab. 13, 149–159 (2011)

NANOTECHNOLOGY

Wired up by DNA strands

In the quest for ever-smaller electronics, DNA could function as a molecular wire, say Jacqueline Barton and her group at the California Institute of Technology in Pasadena. They report that a 34-nanometre-long monolayer of double-stranded DNA can transport electrical charge.

The researchers measured the current of electrons flowing from a gold electrode, down the DNA layer to a probe at the other end. Charge transport required perfect matching between the DNA's base pairs, with just a single mismatch in 100 base pairs hampering electron flow.

This is among the farthest that a molecular wire has transported charge, the authors say. They add that DNA's intrinsic long-range order, flexibility and ease of synthesis make it an attractive molecule for nanoelectronics.

Nature Chem. doi:10.1038/nchem.982 (2011)

FLUID DYNAMICS

What killed the top kill?

Attempts to 'top kill' last year's oil spill in the Gulf of Mexico by ramming mud down the well may have failed because of the unsuitable properties of the drilling mud used.

Using coloured water and mineral oil, Jonathan Katz of Washington University in St Louis, Missouri, and his colleagues show that the outrushing oil may have broken up the incoming column of mud into small

COMMUNITY CHOICE

The most viewed papers in science

PROTEOMICS

Toxic clumps trap many proteins

HIGHLY READ
on www.cell.com
in January 2010

Neurodegenerative disorders such as Parkinson's disease are marked by the presence of toxic protein aggregates in brain cells. These aggregates seem to ensnare a range of proteins that share certain biochemical features and are involved in important cellular processes.

Martin Vabulas at the Max Planck Institute of Biochemistry in Martinsried, Germany, and his colleagues designed artificial β -sheet proteins that form clumps similar to the amyloid fibrils seen in several neurodegenerative diseases. They expressed the proteins in human cells and analysed the native proteins that bound to the aggregates. The researchers found that various proteins involved in RNA processing, protein production and other key functions were caught up in the aggregates: in a few cases, the entangled proportion was as high as 45%. These proteins tended to be larger in size and have more unstructured regions than their non-sequestered counterparts. Moreover, the proteins interact with hundreds of other essential proteins.

Cell 144, 67–78 (2011)

droplets (pictured left, in green). They also show that future wells could be killed more easily by using material that becomes stiff when stretched rapidly — in this study, water with added corn starch (right). This would allow the mud to travel down the well as a slug — without getting churned up — and form a seal.

Phys. Rev. Lett. 106, 058301 (2011)



CANCER GENETICS

Teasing apart cancer's influences

Genetic mutations that arise after birth are one of cancer's main drivers. But what effect do inherited 'germline' genetic variants have on the progression of tumours with

these mutations?

A team led by Allan Balmain at the University of California, San Francisco, studied how germline genetic variants affected the expression of genes associated with skin cancer in mice. The authors used animals genetically prone to developing cancerous tumours that had been treated with tumour-inducing chemicals. They say these mice are more relevant models for human genetic heterogeneity than other mouse strains typically used for cancer research.

The team found that the effect of germline variants on gene expression decreased as skin tumours progressed, indicating that the spontaneous mutations may have rewired genomic networks. However, this rewiring may also have led to the expression of certain genes linked to inflammation and tumour susceptibility coming under germline control in tumours.

Genome Biol. 12, R5 (2011)

NATURE.COM

For the latest research published by Nature visit:

www.nature.com/latestresearch

LAWRENCE LIVERMORE NATL LABORATORY

SEVEN DAYS

The news in brief

POLICY

US budget cuts?

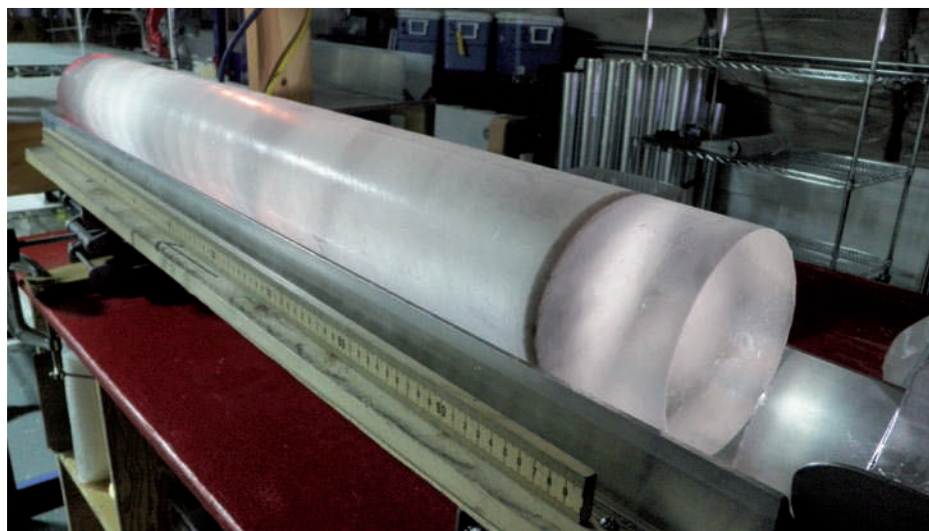
The Republican-controlled US House of Representatives released a budget on 3 February that could mean cuts in federal science funding for the remainder of the 2011 fiscal year (March–September). Among the proposals was a 16% decrease from 2010 levels for the commerce, science, and justice subcommittee, which funds agencies such as NASA and the National Science Foundation. Each subcommittee must now make spending recommendations, which will have to find agreement with the Democrat-controlled Senate and President Barack Obama. See go.nature.com/mdzxm9 for more.

Stem-cell tangle

The legal uncertainty over the status of research using human embryonic stem (ES) cells in the United States is harming work on stem cells in general, according to a survey of 370 researchers released on 3 February (A. D. Levine *Cell Stem Cell* 8, 132–135; 2011). Scientists in the poll said that they were delaying research or moving away from work involving human ES cells. See pages 156–159 for more on the ongoing US lawsuit on the use of such cells.

Clinical research

The US National Institutes of Health (NIH) has launched an elite programme to create a new breed of physician–scientists. The scheme will support three medically trained scholars to conduct clinical research on the NIH's campus in Bethesda, Maryland, for 5–7 years at a cost of around US\$1 million a year each. They will then get up to \$500,000 funding for another 5–6 years at the NIH or elsewhere. The scheme



H. ROOP

Record ice core drilled

Researchers at the West Antarctic Ice Sheet Divide Ice Core project have drilled a column of ice nearly as long as ten Empire State Buildings stacked on top of one another. Gas bubbles in the 3,330-metre-long core, the final section of which was extracted on 28 January, should

provide a 100,000-year climate record. It is the longest ice core ever drilled solely by US scientists, and the second longest ever made. A joint team of Russian, US and French scientists completed the longest ice core, at 3,623 metres, in 1998. See go.nature.com/ffgeg7 for more.

will eventually support 20–30 researchers. See go.nature.com/189ljt for more.

Scientific integrity

The US Department of the Interior laid out a new policy on scientific integrity on 1 February, including a ban on political appointees altering technical findings. The department has been the quickest agency to respond to a March 2009 memo from President Barack Obama that promised to put sound science at the centre of government policy-making. See go.nature.com/jdziwy for more.

Europe united

Heads of the European Union member states have set themselves a deadline of 2014 for completing the European Research Area: a concept

that sees Europe as a unified entity in which researchers and funding can move freely across national borders. Single European patents, portable research grants and transferable pensions are the main sticking points. The agreement was made at a European Council summit on 4 February.

Perchlorate ruling

The US Environmental Protection Agency (EPA) will start to regulate perchlorate in drinking water — a significant moment in a debate that has raged since the late 1990s, when the chemical was discovered in many water supplies. Perchlorate interferes with the production of thyroid hormones and mainly leaches into the environment from its use in the manufacture of

rocket fuel and explosives. Under President George W. Bush, the EPA had decided in 2008 that regulation was not needed. Lisa Jackson, head of the EPA, announced the reversal of that decision on 2 February.

RESEARCH

Retractions rise

A case of scientific misconduct at the Research Center Borstel in Germany is assuming alarming proportions. The centre, which launched an investigation last July, said last week that retractions are under way of 6 further papers produced by current and former members of its immunology group, making a total of 12 withdrawn publications. The head of the group, immunologist

Silvia Bulfone-Paus, says that two former postdocs manipulated images without her knowledge. See go.nature.com/kmyalr for more.

Year of forests

The United Nations (UN) launched its International Year of Forests in New York on 2 February. Marking the event, the Food and Agriculture Organization of the UN in Rome released a biennial assessment of global forests issues. *State of the World's Forests 2011* says that the rate of deforestation has slowed in the past decade, but remains "alarmingly high". The report emphasizes that local communities' knowledge about managing forests should be taken into account in top-down efforts to reduce greenhouse-gas emissions from deforestation.

Arctic fishing

Fishing catches in the seasonally ice-free Arctic Sea by Russia, the United States and Canada were 75 times greater than reported to the United Nations' Food and Agriculture Organization from 1950 to 2006, according to estimates published last week (D. Zeller *et al. Polar Biol.* doi:10.1007/s00300-010-0952-3; 2011). The calculated total over the period, some 950,000 tonnes, is still small. Researchers at the University

of British Columbia in Vancouver, Canada, said that unreported subsistence fishing was mainly responsible.



Amazon pain

According to satellite observations, the drought last year in the Amazon basin (pictured) was even more widespread and intense than the dry spell in 2005, which had been thought to be a once-in-a-century occurrence. If such arid conditions continue, the world's largest rainforest might no longer buffer increases in atmospheric carbon dioxide, wrote researchers on 3 February (S. L. Lewis *et al. Science* 331, 554; 2011).

Medical detectives

An effort to find the causes of mystery illnesses has declared its first success. Researchers at the Undiagnosed Diseases Program at the US National Institutes of Health in Bethesda, Maryland, pinpointed the genetic mutation that causes a rare artery-hardening condition

(C. St Hilaire *et al. N. Engl. J. Med.* 364, 432–442; 2011). See go.nature.com/1naxqr for more.

BUSINESS

Obesity drug upset

US regulators have rejected another obesity drug, despite an earlier recommendation from advisers to conditionally approve it. On 1 February, the Food and Drug Administration (FDA) told Orexigen Therapeutics of La Jolla, California, that concerns about the possible cardiovascular risks of the drug Contrave (naltrexone/bupropion) outweighed its weight-loss benefit. It asked for further clinical trials. Orexigen's share price fell by 72% following the news. Last year, the FDA rejected two other obesity drugs and asked for a third to be pulled off the market.

Research cutback

Pharmaceutical giant Pfizer on 1 February announced cuts to its research budget and the closure of its research centre in Sandwich, UK. Most of the 2,400 staff there are scientists. See pages 141 and 154 for more.

PEOPLE

Developing world

Romain Murenzi, a physicist and Rwanda's former science minister, was named on

COMING UP

14 FEBRUARY

NASA's Stardust mission — rebranded NExT — is due to fly by the comet Tempel 1. It is the first follow-up mission to a comet: the Deep Impact mission targeted Tempel 1 five years ago. go.nature.com/1ho7bt

14 FEBRUARY

US President Barack Obama submits his 2012 budget request.

17–21 FEBRUARY

The American Association for the Advancement of Science holds its annual meeting in Washington DC. www.aaas.org

7 February as the new executive director of TWAS, the academy of sciences for the developing world. Murenzi is expected to take up the post at TWAS headquarters in Trieste, Italy, around April; he will replace Mohamed Hassan, who has spent 25 years in the post. See go.nature.com/jf2zct for more.

ALS prize

American neurologist Seward Rutkove has won a US\$1-million prize for creating a non-invasive tool that tracks the progress of the neurodegenerative disease amyotrophic lateral sclerosis (ALS). The biomarker developed by Rutkove, of Beth Israel Deaconess Medical Center in Boston, Massachusetts, detects diseased muscle tissue by sending electrical currents through the body. The ALS biomarker award was launched in 2006 by Prize4Life, a foundation based in Cambridge, Massachusetts, to spur breakthroughs in treating the disease.

➔ NATURE.COM

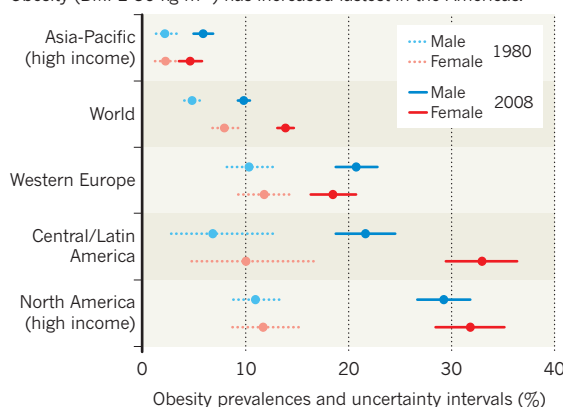
For daily news updates see: www.nature.com/news

TREND WATCH

Obesity rates worldwide almost doubled between 1980 and 2008, an analysis of health-examination surveys has found (M. M. Finucane *et al. Lancet* doi:10.1016/S0140-6736(10)62037-5; 2011). In 2008, 9.8% of men and 13.8% of women were obese, as measured by a body-mass index (BMI) of at least 30 (kilograms weight per square metre of height). Rates of obesity were highest in men in North America and women in southern Africa, and lowest in south Asia for both men and women.

THE WORLD GAINS WEIGHT

Obesity (BMI ≥ 30 kg m⁻²) has increased fastest in the Americas.



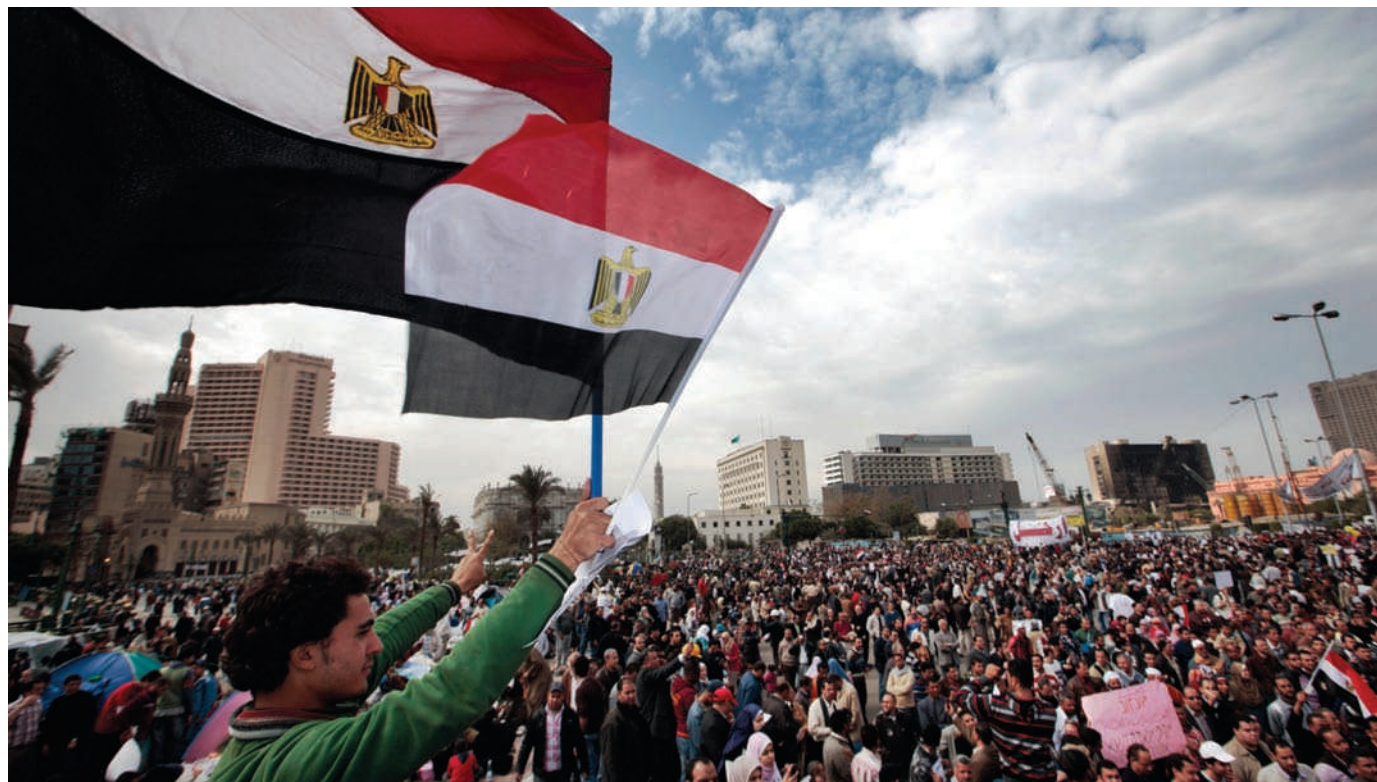
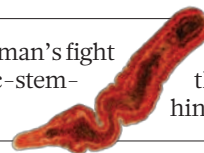
NEWS IN FOCUS

COMMUNITY One year after a massacre, a shattered biology department rebuilds **p.150**

GENOMICS Single-molecule DNA sequencing proves to be fast but error-prone **p.154**

BIOMEDICINE One woman's fight against embryonic-stem-cell research **p.156**

EVOLUTION A battle over the ancestry of animals hinges on a lowly worm **p.161**



AP/L. PITRAKIS

Thousands of anti-government protesters, including many students and academics, rally in Cairo's Tahrir Square.

POLICY

Egypt's youth 'key to revival'

Country's future depends on democracy, education and research reform, say scientists.

BY DECLAN BUTLER

There is a pool of talent among the youth in Egypt that is unbelievable. They are people who think creatively and critically, who are yearning for the freedom to express themselves, and many of them are those who are leading this revolution." Tarek Khalil, president and provost of the non-profit, independent Nile University in Cairo, is convinced that the country's Facebook generation represents the best chance in decades for the intellectual renaissance of a society that has been rendered moribund and impoverished by the military dictatorship of President Hosni Mubarak.

Other Egyptian researchers contacted by *Nature* share Khalil's views (see 'Scientists speak out'). They emphasize that, with the regime still in place two weeks after anti-Mubarak protests began on 25 January, the most urgent priorities are to halt the regime's crackdowns on protesters, and to ensure that the pro-democracy movement prevails. But in the long term, they say, Egypt's education and science systems must be completely overhauled to help address the root causes of its social and economic woes.

"The current outdated government simply lacks the mindset and vision to strategically support scientific research and lead an innovation-based economy that can compete globally,"

says Hassan Azzazy, a chemist at the non-profit, private American University in Cairo.

In an editorial in the *International Herald Tribune* last week, Ahmed Zewail, an Egyptian-born researcher at the California Institute of Technology in Pasadena, who won the 1999 Nobel chemistry prize, slammed the regime for presiding over a long deterioration in Egypt's education and research systems. Zewail returned to his home country last week to join a group of prominent Egyptian intellectuals who are drawing up plans, including constitutional reforms, to try to engineer a peaceful transition to democracy. Last week, Zewail — who is also one of six science envoys appointed by US President Barack Obama ►

COLLABORATION

Synchrotron faces setback

The Synchrotron-light for Experimental Science and Applications in the Middle East (SESAME) project began in the late 1990s with a dual aim: to do research while building scientific ties in the troubled region. The plan is to install and upgrade a decommissioned German synchrotron at a facility in Alaan, Jordan, that scientists from partner nations will use for materials science and biological imaging.

But the project needs an extra US\$35 million to complete construction (on top of the \$55 million to \$60 million already committed), and it was counting on Egypt. Last year, Israel promised \$5 million, on the condition that other partner nations put in similar amounts. Egypt was among those expected to match the Israeli pledge at a SESAME meeting on 11 March. Hany Helal, the nation's science minister under President Hosni Mubarak, has been a staunch supporter of SESAME, but as *Nature* went to press, it was unclear how long Helal would remain in his post, or how a new government might view the project.

"It's obviously a bit worrying," says Chris Llewellyn Smith, a British physicist and president of SESAME's council. "But I think we'll come through it." Indeed, scientists across the Middle East are adamant that SESAME will proceed despite the unrest in Egypt, anti-government protests in Jordan, and the murders of two members of the project's Iranian delegation (see *Nature* 468, 607; 2010). "It's very important that we keep it going, especially at times like this," says Zehra Sayers, a biophysicist at Sabanci University in Istanbul, Turkey, who chairs SESAME's science advisory committee.

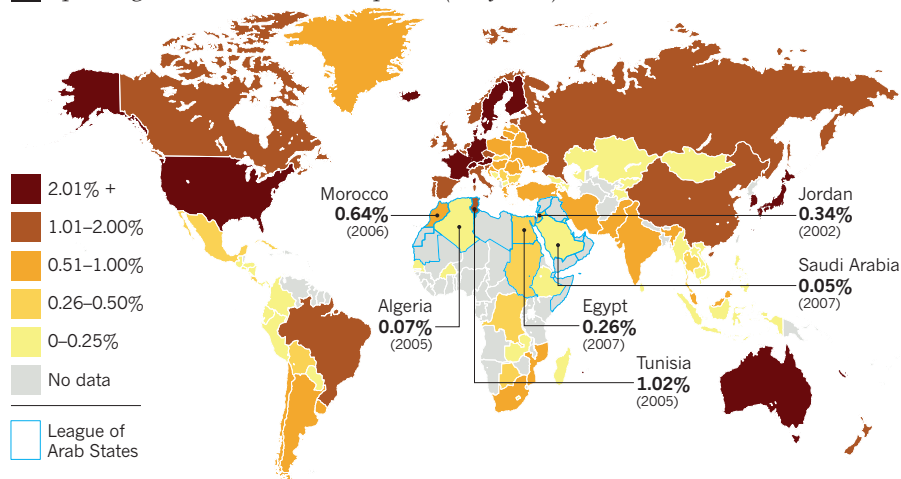
Jordan, Iran, Turkey and the Palestinian Authority had all indicated that they might match Israel's offer, and Llewellyn Smith says that even if funding falls short, it could spur extra support from Western governments and major foundations.

Tarek Hussein, a physicist at Cairo University who has been encouraging his students to take part in the peaceful protests, says he is optimistic that any new government will remain committed to SESAME. There is reason for hope: Mohammed ElBaradei, a leading Egyptian opposition figure, was supportive of SESAME when he was director of the International Atomic Energy Agency in Vienna. **Geoff Brumfiel**
See go.nature.com/pl2jil for a longer version of this story.

STAGNATING SCIENCE

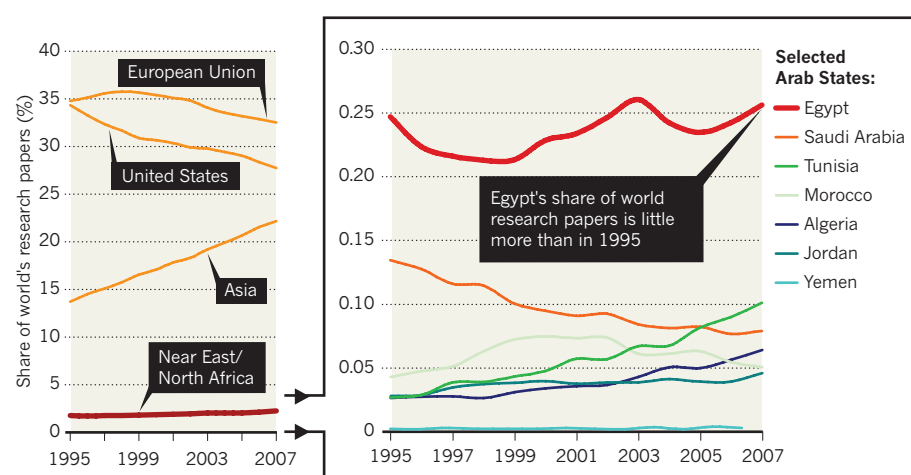
Science is poorly funded in Arab states. In 2007, they spent on average only 0.3% of their GDP on research, compared with a world average of 1.7% (1). Egypt leads the region on research publications, but its world share has remained flat for more than a decade (2).

1 Spending on research and development (% of GDP)



SOURCE: UNESCO

2 Share of science publications



SOURCE: NSF SCIENCE INDICATORS 2010, SCI

► to Muslim-majority countries — called on Mubarak to step aside, to help Egypt make a fresh start.

Zewail's assessment of Egypt's decrepit education and research systems is accurate, says Khalil. Intake at the public universities — which offer students free tuition — has expanded vastly since the 1960s, in line with the country's rapidly growing population. But budgets have remained flat, salaries have stagnated, and training of teachers and lecturers has been neglected, he explains. "Egyptian public universities currently do not foster productivity or innovation," adds Azzazy. "They are simply assembly lines that produce thousands of unskilled graduates every year."

As is the case in other authoritarian Arab regimes (see *Nature* 469, 453–454; 2011), political patronage and nepotism are rife in senior university appointments. The suppression of human rights, and the poor conditions

for science, have also led to a brain drain to the West, and more recently to Gulf states that are investing in research. According to the Science Citation Index, Egypt produced 5,140 scientific papers in 2010. Harvard University in Cambridge, Massachusetts, published twice that number alone.

Egypt's research has also been plagued by a lack of funding, with research spending amounting to just 0.2–0.3% of the country's gross domestic product (GDP; see graphic). "Mubarak wasn't interested in science; he didn't have the vision or the ability to understand what development takes," says Khalil. "Hopefully, we will get an Egypt that will appreciate research and education. This has to be a top priority for the country."

Fostering international collaborations could be a key factor in achieving

NATURE.COM
For interviews with Hassan Azzazy and Tarek Khalil, see: go.nature.com/nvzvbk and go.nature.com/kuonbt

ANTIQUITIES

The fight to protect an ancient heritage

Images of Molotov cocktails thrown towards the Egyptian Museum in Cairo by supporters of President Hosni Mubarak last week horrified archaeologists, who feared for the world's largest collection of Egyptian artefacts. Vandalism on 28 January had already damaged about 70 pieces there.

Such attacks highlight the predicament of protecting cultural heritage during conflicts, while also dealing with a more important priority: preventing loss of life. Under the 1954 Convention for the Protection of Cultural Property in the Event of Armed Conflict, countries pledge to create inventories of heritage areas, and to set up military units with archaeological expertise to protect sites. Egypt has signed the treaty, but the unrest caught its military off guard — protesters and citizens took the lead in protecting museums and sites.

Zahi Hawass, the newly appointed Egyptian minister of antiquities, said last week that the army was now protecting all 24 national museums and all major archaeological sites. But some fear that may not be enough. "Egypt's ancient heritage is so rich that the whole country is basically one large open-air museum. It would be impossible to station a soldier at the door of each and every tomb," says Margaret Maitland, an Egyptologist at the University of Oxford, UK, who has been writing about the incidents on her blog, *The Eloquent Peasant*.

Frank Rühli, co-head of the Swiss Mummy Project at the University of Zurich, hopes Egypt will ask external experts to assess the



A damaged mummy at the Egyptian Museum.

damage and decide on restoration measures, as was done in 2003 during the Iraq War by the United Nations Educational, Scientific and Cultural Organization. Researchers have asked law-enforcement agencies and art dealers around the world to look out for stolen antiquities.

Yet Egypt's heritage is understandably not foremost in the minds of those facing the violent crackdown. "We are all very concerned about the Egyptian Museum, but please what we need first is to restore order and save the Egyptian people," said a member of the Restore + Save the Egyptian Museum! Facebook page last week, after bloody clashes in Cairo's Tahrir Square. **D.B.** See go.nature.com/v4fbui for a longer version of this story.

that. But the unrest has already led to uncertainty about Egypt's role in the SESAME synchrotron project in Jordan (see 'Synchrotron faces setback'), and looting at the Egyptian Museum in Cairo and other cultural-heritage sites has raised archaeologists' fears over the security of the country's antiquities and possible threats to research (see 'The fight to protect an ancient heritage').

But it is the denial of freedoms under the Mubarak regime that Egyptian scientists see as the most serious obstacle to progress. The stifling effect that this has had on creativity is "detrimental to creating a knowledge society; people dare not think out-of-the-box", says Farouk El-Baz, an Egyptian-born geologist at Boston University in Massachusetts.

One result, says Khalil, is that the educational culture in Egypt has become based on rote learning of existing knowledge and dogma, and doesn't allow for debate or creative thinking. "The whole concept of creativity

and entrepreneurship is alien to the existing system," he says.

"Building science is not just a question of money and projects, it is also about a whole climate of research, of freedom of enquiry, freedom of expression, education, the ability to question," adds Ismail Serageldin, director of the Library of Alexandria. That the country's youth is now standing up for these values gives reason for hope, he says.

Despite the repression and stagnation in Egypt, Serageldin says, profound changes have been brewing for years. Empowered by discussions using the Internet, the young have come to find freedom of expression, and other rights, "so natural that it's like breathing — they can't accept anything else", he says.

"What Egypt most needs now to develop itself is to unleash the energy of its youth and its population," adds El-Baz. "This regime must leave, and let a younger generation take power." ■

Scientists speak out

Several senior academics took to the streets of Cairo last week to join the anti-government protests. They spoke to Nature Middle East about the state of science in Egypt and their hopes for the future.

Mahmoud Saleh Chemist, Cairo University

There is no proper scientific research in Egypt. This regime has killed the talents and capabilities of the Egyptian people, whether scientific, social or political.

Spending on scientific research was minimal. A whole generation of scientists moved overseas to continue their pursuit of knowledge to help this country. But when they returned they found that all paths to achieve this were blocked. This is part of the reason behind us revolting.

We hope there will be a new, smarter government that respects the role of science and technology in the development of society.

Hani Dewedar Astronomer, Cairo University

University education needs serious reform. The government needs to invest and provide a good environment for education to maximize the student experience.

We want to link universities to the community and to industry. This would be of great benefit to the students and the education system as well as society in general.

Tahir Ahmed Yehia Agricultural scientist, Cairo University

Scientists need tools: funding, proper equipment, good budgets for universities and laboratories. Sadly, however, these tools are lacking in Egypt and it is apparent that the current regime does not believe in scientific research.

In this protest there is no distinction between university professors and students. We've all come out as Egyptians. There is no distinct age or social standard, we all have the same demands — a regime change that improves conditions for us all and tackles the problems we have faced for so long.

INTERVIEWS BY MOHAMMED YAHIA
See go.nature.com/gbd12q for longer versions of these interviews.



B. DILL/NATURE

Joseph Leahy (centre), critically wounded a year ago, is back on campus nearly every day.

COMMUNITY

University seeks to emerge from shooting's shadow

One year after an assistant professor murdered three colleagues at the University of Alabama in Huntsville, researchers are striving for a new 'normal'.

BY MEREDITH WADMAN

Under normal circumstances it would have been a year of modest successes for a small but ambitious biology department. A new confocal microscope is up and running; spring enrolment is up, four PhD candidates are expected to graduate and, last month, the Carnegie Foundation, which classifies institutions of higher education in the United States, changed its rating of research activity on campus from 'high' to 'very high'.

But for the biology department at the University of Alabama in Huntsville (UAH), such achievements stand in a more heroic light. Just one year ago, on 12 February 2010, biologist Amy Bishop opened fire with a 9-millimetre pistol during a faculty meeting, killing three department members and critically wounding two others (see *Nature* **465**, 150–155; 2010). A third sustained a less-serious injury. Now, the department is labouring to regain its balance, coping with the logistics of rebuilding and with the legal fallout from the tragedy.

On Saturday, members of the UAH community will gather to mark the anniversary

of the shooting with a service of 'remembrance and renewal'. Students and faculty members will speak about the legacy of the fallen: plant biologist Maria Ragland Davis; physiologist Adriel Johnson; and Gopi Podila, a plant molecular biologist who was chairman of the biology department. They will also pay tribute to the survivors: microbiologist Joseph Leahy and staff assistant Stephanie Monticciolo, both still recovering from grievous injuries. Others — including the rest of the biology faculty, nearly all of whom were gathered around the table in the small conference room where the shootings occurred — nurse wounds that are not as visible.

The university is at pains to note that the service will look forwards as much as back. "The event is part of our history. We cannot take it away," UAH president David Williams told *Nature* last week. "But we will not let it define the university."

For the 12-member biology faculty, the goal "is to be at a point where we are recognized not as a department of tragedy,

but as a department of academic excellence", says Joseph Ng, a structural biologist who witnessed the shootings.

Ng, who is leading the search to permanently fill the vacant positions, points to signs of progress. As of last week, he had received nearly 200 applications for three tenure-track positions that will help to restore the department to full strength. For now, visiting professors are teaching the physiology, cell biology and microbiology that had been taught by Bishop, Davis, Johnson and Leahy. Colleagues are shepherding the last months of funding from grants that belonged to the slain biologists. The new confocal microscope for which Podila and Davis had won funding is in use. And Debra Moriarity, a biologist who had been graduate dean at UAH, has taken over as interim chair of the department, after a collective decision to delay the search for a permanent chair. Moriarity, who crawled under the conference table and grabbed Bishop by the leg to try to stop her shooting, says that the toughest thing she has had to do in recent months is to move into Podila's office, packing up his stuff. "We are trying to figure out what the new 'normal' is," she says.

NATURE.COM
The aftermath
of the Huntsville
massacre:
go.nature.com/qitvmy

Although Bishop has not yet been indicted for the shootings, a violent personal history has emerged that has continued to cast a shadow over UAH's efforts to recover. In June, a Massachusetts grand jury charged Bishop with fatally shooting her brother in 1986, an act that had originally been ruled an accident. Last month, the spouses of Johnson and Davis filed a 'wrongful death' lawsuit against Bishop, her husband James Anderson and UAH provost Vistasp Karbhari. It alleges that Karbhari was "directly aware of Bishop's emotional instability" and failed to implement a university safety policy that would have required contacting police or counselling services. (Leahy, Monticciolo and their spouses filed a suit against Bishop and Anderson, without naming Karbhari, in November.)

The university has responded in a statement that it "will vigorously defend" the lawsuit. It adds that it "is saddened by the decision to sue Dr. Vistasp Karbhari and does not agree that Dr. Karbhari, or anyone associated with the university, could have predicted or prevented this random act of violence".

The return of Leahy has been an inspiration to faculty members and students. Bishop fired her last bullet at him before her gun jammed, shattering his forehead and severing his right optic nerve; today he is blind in his right eye and has partial vision in his left. The bullet remains

lodged in his neck, too close to the jugular vein to remove safely. A metal plate inserted to replace part of his skull became infected last autumn and had to be removed, leaving a baseball-sized indentation in his forehead. Another operation will be required to insert a new plate. Until that happens, falling and injuring the underlying brain remains a significant risk.

"You come to grips with the fact that it has happened. You just want to make sure that it doesn't defeat you."

runs at the fitness centre; works in his office, where an extra-bright lamp has been installed; and assists with a class being taught by Moriarity for aspiring health-profession students, whom he has also begun advising — a job that had been Johnson's. He has tentative okays from Moriarity and the dean of science to return to work full time in the autumn.

Not for a moment did he think of leaving biology, says Leahy, an expert in the bacterial degradation of hydrocarbons. "This is who I am. This is what I do."

Nonetheless, Leahy, who has no memory of the shooting, is now on campus almost every day, when he is not at one of the innumerable therapy sessions — speech, vision, physical and occupational — that have consumed the past year. He lifts weights or

"I find him incredibly inspirational: 'Yes, I have a huge crater in my head, but it is what it is and life goes on,'" says Leland Cseke, a research professor who saw Bishop rush down the hall as she fled after the shootings. "When you see him, you just have a good feeling."

Bishop, in the meantime, awaits her fate in the Madison County jail. She stands charged with one count of capital murder — murdering two or more people at the same time is a capital offence in Alabama — and three counts of attempted murder. A grand jury is likely to hear the facts of the case in the coming weeks. In Alabama, capital murder is punishable by life in prison without parole or by the death penalty. Robert Becher, the chief trial attorney for the Madison County District Attorney's Office, who is prosecuting the case, says that the state has not yet decided which punishment to seek.

Back on campus, another decision is pending: what to do with the conference room where the shootings occurred, which remains closed and locked. It may be converted into an open gathering space. Or it may be reconstituted as a conference room, says Moriarity, "so it doesn't look the same but could still be used — with two doors".

Moving forward, with a focus on rebuilding, is vital, says Ng. "You come to grips with the fact that it has happened. You just want to make sure that it doesn't defeat you." ■

DISCRIMINATION

Science gender gap probed

Overt sexism is no longer the norm, but societal barriers remain for women in science.

BY GWYNETH DICKEY ZAKAIB

Goodbye glass ceiling; so long old-boys club. The metaphor that best describes the challenge facing women in science today is the invisible web. Its multiple strands — some social, some biological, some institutional — can make it significantly harder for female researchers to achieve as much, as fast, as their male counterparts.

So concludes a study that set out to explore the persistent gap in the number of women in maths-intensive fields such as physics, computer science and engineering. It finds that overt discrimination of the sort that would make a female candidate less likely to be hired, published or funded when competing against an equally qualified male is largely a thing of the past. Instead, trade-offs between pursuing a career and raising a family, coupled with societal factors and gender expectations that can influence professional choices at a young age, are more likely to account for the shortage of women in some fields.

A 2008 survey of US universities by the National Science Foundation revealed that less than 30% of PhDs in the physical sciences were awarded to women. Higher up the ranks, women make up only about 10% of full professorships in physics-related disciplines. Yet when psychologists Stephen Ceci and Wendy Williams of Cornell University in Ithaca, New York, sifted through 20 years of research, they found little evidence of continued gender bias in journal reviewers, granting agencies or hiring committees. Their analysis, published on 7 February (S. J. Ceci and W. M. Williams *Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.1014871108; 2011), contrasts with reports that suggest overt discrimination remains a significant problem.

"There are constant and unsupportable allegations that women suffer discrimination in

these arenas, and we show conclusively that women do not," says Williams.

Ceci and Williams conclude that female researchers lag behind their male counterparts in professional advancement because of a broader set of societal realities. Much of the problem, they say, can be boiled down to external factors related to family formation and child rearing. Motherhood can make women less likely to choose research careers than male scientists of equal ability, or lead them to

academic disciplines, Ceci and Williams point out, so additional factors must account for why the gender gap in science is greatest in maths-intensive fields. This could include a difference in the fields women prefer, a choice that can be socially influenced, regardless of aptitude.

"They are probably right that overt discrimination has declined, but it's naive to suggest that judging applicants differently based on their gender is a thing of the past," says Christianne Corbett, a senior researcher at the

American Association of University Women (AAUW), based in Washington DC. In their study, Ceci and Williams criticize an AAUW report published last year that claims there is discrimination in peer review and "a systematic under-rating of female applicants" in hiring.

Ceci and Williams say that a continuing focus on discrimination could be drawing attention away from the true causes of the disparity. They suggest, for example, that gender-sensitivity training for review and hiring committees may no longer be needed, and argue that efforts should be redirected to promoting flexible tenure policies for women with young children. Educational programmes could also help female graduate students to make more informed decisions about family and career.

Nancy Hopkins, a molecular biologist who chaired a landmark study of gender inequality in faculty members at the Massachusetts Institute of Technology (MIT) in Cambridge, cautions that progress in gender equality could backslide if successful practices are abandoned.

She observes that some family-friendly changes are already being made at MIT, such as the availability of on-campus day care for faculty members with young children. "We're about two-thirds of the way home," Hopkins says, but she notes that many institutions have further to go. ■



Too rare a sight: working on a neutrino experiment at Fermilab.

choose academic positions with larger teaching loads but more regular hours, sacrificing time for research. The authors also point out that the strict tenure timeline conflicts directly with women's window for child rearing.

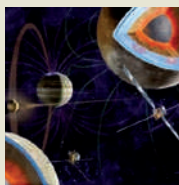
"A woman who has young children is still expected to come up for tenure 5–6 years after she starts her job," says Williams. "It creates a virtually insurmountable obstacle."

Such constraints affect women across all

P. GINTER/SCIENCE FRACTION/CORBIS


**MORE
ONLINE**

LATEST NEWS



ESA mission selection has consequences for international collaborations
go.nature.com/ilu2m9

MORE NEWS

- Are antipsychotics shrinking patients' brains? go.nature.com/hteezz
- Outdoor mosquitoes could defy control go.nature.com/7kzztq
- Fly brain structure illuminated go.nature.com/ubwbuj
- Solar-sail success go.nature.com/y6Gapod

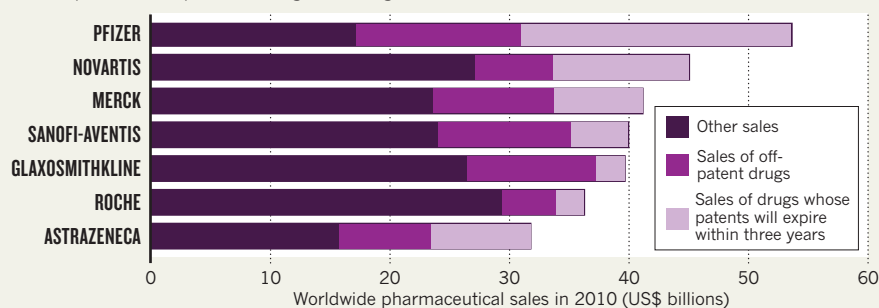
ON THE BLOG



Citizen scientists beat Kepler team to likely planets
go.nature.com/1zgra9

OVER THE CLIFF

As pharmaceutical companies experience a slew of patent expiries, revenues will become much more exposed to competition from generic drugs.



ESTIMATED FIGURES FROM EVALUATEPHARMA

PHARMACEUTICALS

Pfizer slashes R&D

Drug-maker plans to cut jobs and spending as industry shies away from drug discovery.

BY DANIEL CRESSEY

The pharmaceutical industry has spent years bracing itself for the 'patent cliff', when sales are expected to plummet as a bundle of blockbuster drugs loses protection against generic competitors.

Yet the dire consequences for drug research — and the scientists behind it — still took many by surprise last week. With key patents about to expire, Pfizer, the world's largest pharmaceutical company in terms of sales, unveiled plans to slash its research and development (R&D) spending by billions and cut thousands of jobs.

The chief casualty is the company's research facility in Sandwich, UK, where the erectile dysfunction drug Viagra was developed. The site will close in 18–24 months, with almost all of the 2,400 employees there — mostly scientists — facing redundancy. Meanwhile, the company's R&D headquarters in Groton, Connecticut, will be shedding roughly 1,100 jobs. Pfizer will also cut R&D expenditure for 2012 to between US\$6.5 billion and \$7.0 billion, down from its previous target of \$8 billion to \$8.5 billion.

"There are going to be a lot of unemployed scientists," says Ashley Woodcock, head of the School of Translational Medicine at the University of Manchester, UK. He fears that Pfizer's cuts, on top of job culls announced by GlaxoSmithKline and AstraZeneca last year, will erode traditionally strong links between the drug industry and academia in Britain. "Having industry alongside academic research is what has made us great," says Woodcock, "and if the industry leaks away we'll be in real trouble."

Pfizer plans to ditch research into areas including allergy and respiratory diseases, which are based at its Sandwich site, although

it is unlikely to completely abandon promising candidate drugs. The company plans to focus on neuroscience, oncology, vaccines, and cardiovascular and inflammation treatments.

This may not be the company's last retrench. London-based analysts EvaluatePharma expect patent expiries over the next three years to expose about two-thirds of Pfizer's total sales to competition from generics (see graph). This is largely a consequence of the company's reliance on its cholesterol-reducing drug Lipitor — the world's best-selling drug last year — which will lose patent protection in November.

Rival companies also face sales losses running into the billions. The latest parade of annual results shows that although profits are buoyant for now, companies are increasingly reluctant to sink money into R&D pipelines, which have been slow to yield new drugs. Instead, share buy-backs — which boost share prices and please investors — and outsourcing R&D are in vogue.

GlaxoSmithKline last week said it would be spending between £1 billion (US\$1.6 billion) and £2 billion in 2011 as part of a "long-term share buy-back programme", and added that it has made "fundamental changes to how we allocate our R&D expenditure". The company plans to focus its work on getting the most promising drugs to market and will cut costs and risk "through externalising parts of early-stage discovery; dismantling infrastructure; and terminating development in areas with low financial and scientific return".

"The pharma industry is deciding its core capabilities are marketing and dealing with regulatory bodies," says Judy Slinn, a business historian at Oxford Brooks University, UK. "Pharma companies will still do development work. They won't do discovery." ■ SEE COLUMN P141

GENOMICS

Gene reading steps up a gear

Third-generation sequencing machines promise to make their mark one molecule at a time.

BY HEIDI LEDFORD

It's super cool, but it's never going to work," genomics guru Eric Schadt responded when a wary investor asked for his opinion about a new DNA-sequencing technology in 2003. A company was creating a machine that it claimed could revolutionize the field by reading over the shoulder of an enzyme as it copied DNA molecules.

Despite his initial scepticism, Schadt touted the method's success last weekend at the Advances in Genome Biology and Technology meeting in Marco Island, Florida. Now chief scientific officer at the company he had once doubted — Pacific Biosciences in Menlo Park, California — Schadt was one of several researchers at the meeting who provided a glimpse of how the company's first DNA-sequencing machines are performing.

All eyes are on these machines. Pacific Biosciences set a high bar for its own success in 2008, when chief technology officer Stephen Turner boasted that the instruments would be able to sequence a human genome in just 15 minutes by 2013, compared with the full month it took at that time. This year, as researchers unveiled data from the first machines to leave the company's campus, the discussion was less about revolutionizing the field and more about niche applications.

After several delays, customers have now been told to expect their machines in the second quarter of this year. The machines potentially offer advantages over the 'next-generation' sequencers currently on the market. Users of the new machines last week reported generating sequences an average of 1,500 base pairs long — about ten times the length of those currently produced by the state-of-the-art sequencers from Illumina in San Diego, California. These longer reads make it easier to stitch fragments of DNA sequences together into a coherent genome sequence.

Pacific Biosciences' machines are also fast. In a paper published online in December, Schadt and his team used them to trace the origin of the ongoing cholera outbreak in Haiti by sequencing the genomes of five strains of *Vibrio cholerae* (C. S. Chin *et al.* *N. Engl. J. Med.* **364**, 33–42; 2011). The team sequenced all five strains in less than an hour. It takes about a week to complete a 150-base sequencing run

"Single molecule is the future of sequencing, but it still has hurdles."

IN A FLASH

New DNA sequencers watch an enzyme called DNA polymerase as it uses fluorescently tagged bases to synthesize DNA. Each base is identified by a distinguishing colour that flashes as the base is incorporated into the DNA strand.



on an Illumina sequencer.

But for many researchers, the key advance of the Pacific Biosciences machines is the ability to sequence single molecules of DNA. The instruments work by watching as an enzyme confined within a tiny compartment copies DNA, adding fluorescently labelled bases that flash with characteristic colour as they are added to the DNA strand (see 'In a flash'). Leading sequencers on the market instead report an average sequence taken from a population of molecules.

Single-molecule sequencing opens the door to analysing rare sequence variants, and frees researchers from having to amplify DNA samples before sequencing — a step that can introduce errors, and can fail altogether for certain DNA sequences. "Single molecule is the future of sequencing," says Michael Metzker, who studies sequencing technology at Baylor College of Medicine in Houston, Texas. "But it still has hurdles."

Chief among those hurdles has been high error rates. Whereas other methods on the market surpass 99% accuracy, users of the Pacific Biosciences machines last week reported an accuracy rate of about 85%. Schadt argues that this can be overcome by resequencing the same molecule repeatedly.

Nevertheless, because of the cost of its machines (US\$700,000 per unit compared with less than \$125,000 for the new Illumina sequencer rolling out this autumn) and limits on the number of sequences that can be read during every run, the instruments are unlikely to disrupt the sequencing market in the near future. For now, the machines are likely to be used for tackling regions of the

human genome that resisted conventional sequencing. The instruments can also detect some chemical modifications to DNA, which could be useful to the burgeoning epigenetics field. Peter White, who heads the sequencing centre at Nationwide Children's Hospital in Columbus, Ohio, says he is interested in acquiring a machine, but would mainly use it to analyse microbial genomes, which tend to be much smaller than mammalian genomes.

At the meeting last week, Turner did not reiterate his pledge for a 15-minute human genome. But he did emphasize that there is still plenty of room for the current instrument to improve. "We are just at the beginning of this technology." ■

CORRECTIONS

The News story 'Social science lines up its biggest challenges' (*Nature* **470**, 18–19; 2011) should have said that Nick Nash did his MBA at Stanford University.

The News Feature 'Exoplanets on the cheap' (*Nature* **470**, 27–29; 2011) should have said that the spectrometer on which the comb at the Hobby-Eberly Telescope was mounted came from Pennsylvania State University not the University of Pennsylvania.

The graph in the News Feature 'The End of the Wild' (*Nature* **469**, 150–152; 2011) showing a correlation between rising minimum temperatures in Wyoming and increased survival rates for mountain pine beetles should have made it clear that the beetle data were modelled not measured.

➔ **NATURE.COM**
See our human genome special at:
go.nature.com/ugle41



THE CRUSADER

Theresa Deisher once shunned religion for science. Now, with renewed faith, she is fighting human-embryonic-stem-cell research in court.

BY MEREDITH WADMAN

Theresa Deisher was 17 years old the first time she saw a human fetus. Having graduated from the Holy Names Academy in Seattle, Washington, in 1980, she had taken a summer job in the pathology lab at the city's Swedish Hospital when a friend and co-worker miscarried in her fifth month of pregnancy. The fetus arrived fixed in formalin, and Deisher helped to section it to determine the cause of the miscarriage. The body hardly seemed to be the remains of a sentient, soul-bearing human, as the faith of her upbringing had taught, recalls Deisher. Instead, "It looked like a space alien," she says. "I called it 'the thing' for so many years."

Thirty years later, Deisher sees the unborn in a different light. She has reversed her views on embryos and become one of two plaintiffs in a lawsuit filed in 2009, seeking to stop the US government from funding human-embryonic-stem-cell research. The courts hearing the case could issue a decision at any time; many, including Deisher, expect that the matter will end up before the US Supreme Court.

Deisher's co-plaintiff, James Sherley, an adult-stem-cell scientist at the Boston Biomedical Research Institute in Watertown, Massachusetts, is well known as a provocateur. In 2007, he went on a hunger strike to protest against a decision by the Massachusetts Institute of Technology (MIT) in Cambridge to deny him tenure, which he attributed to racism.

Deisher is less well known. A cellular physiologist educated at Stanford University in Palo Alto, California, she spent 17 years in the biotech industry at companies including Genentech, Immunex and Amgen. Three years ago, she founded a tiny, privately held Seattle firm called AVM Biotechnology — the name is a loose abbreviation for 'Ave Maria' — which is dedicated to hastening adult-stem-cell therapies to the market, and to developing alternatives to vaccines and therapeutics made using cell lines from aborted fetuses. She has also launched a non-profit group, the Sound Choice Pharmaceutical Institute, which among other things is investigating, as she puts it, "the potential link between human DNA in childhood vaccines and autism".

Deisher, who is 48 and goes by the name Tracy, is smart, driven and committed. A devout Catholic and a divorced mother of two boys aged 9 and 12, she rises as early as 3:45 a.m. to ride an exercise bike while praying the rosary. She is casual and unpretentious, with a dry humour and a can-do attitude: she spent New Year's Eve laying carpet in the 180-square-metre office space that her company recently moved into.

She is also a bundle of contradictions: an adamant right-to-lifer, whose closest, long-standing friends are pro-choice liberals. She made a healthy six-figure salary at the cream of US biotech companies, but thought nothing of mortgaging it all to launch a no-name firm as the economy slid into a recession. She is a no-frills dresser who has worn a simple gold cross virtually every day for the past 18 years. But she flaunts her intellect. In the past, she alienated friends with a formidable vocabulary fed by a dictionary-reading habit. And she says that those at her church who disagree with her stem-cell views "oftentimes need some education".

Above all, Deisher is supremely confident in her positions, including her attempt to prevent hundreds of millions of dollars from going to human-embryonic-stem-cell research. "It's very difficult to get passionately, morally protective of what physically truly is a clump of cells," she says. "But that is a human being. Scientifically, you can't debate that."

Her arguments, now part of a national discussion, can be hyperbolic. And she does not shy away from assigning motivations to her ideological foes. She says, for example, that embryonic-stem-cell scientists are mostly attracted to the cells' convenience — their rapid growth and what she calls the ease of working with them in the lab. Their science, she says, "is not about helping patients and it's not about advancing the common good".

"I wish that Tracy weren't so polarizing," says Chuck Murry, co-director of the Institute for Stem Cell and Regenerative Medicine at the University of Washington in Seattle, who has known Deisher since they were postdocs together at the university in the early 1990s. "She's kind of the Sarah Palin of stem cells. It would be so much easier to have more rational discourse rather than somebody who heats up the vitriol like this." Deisher counters that she sticks to scientific arguments: "My approach to the stem-cell issue is to remove the polarizing moral debates and speak and educate only about the science."

REGAINING THE FAITH

Deisher showed a bent for science early, teaching herself calculus to win a state competition in which high-school students had to plot the orbit of Mars and design a spaceship and flight path to get there. "Tracy was always very much a leader, an independent thinker," says Liz Swift, who taught Deisher physics at Holy Names and is now the school's principal. In those days, a fun Friday night for Deisher meant several hours at the University of Washington's astrophysics laboratory, followed at 10 p.m. by an outing with girlfriends — only after her mother had checked her for make-up and low necklines.

As a girl, Deisher was torn between her mother's conviction that life began at conception and the views of her two outspoken aunts, both staunch supporters of Planned Parenthood, who reminded her regularly: "It's not a baby. It's a clump of cells." Deisher's experience as a teenager in the Swedish Hospital pathology lab left her without any doubts as to who was right. "I walked out of that lab that weekend and I threw my faith in the garbage can," she recalls.

Weeks after her experience with the fetus, Deisher began undergraduate studies at Stanford, where she went on to earn her PhD in molecular and cellular physiology. On the side, she worked at Genentech in South San Francisco, California, developing assays to support the company's anti-platelet agents. "I was very left-wing," she says. "I was in science, and science was much more interesting than religion. I encouraged a couple of friends to have abortions," urging them to trust her first-hand experience with a fetus in formalin.

Several years later, during an anatomy lab, she encountered the cadaver of a woman also embedded in formalin — looking, she says, not so very different from "the thing". It suddenly struck her that the fetus's 'alien' looks may have simply been attributable to the preservation process. That opened up what she calls "a long, slow process" of coming back to the faith of her childhood. It was one of three pivotal experiences that she talks about as having influenced her decision to actively fight against embryonic-stem-cell research.

After completing a postdoctoral fellowship at the University of Washington in 1993, Deisher went to work for the biotech company Repligen in Waltham, Massachusetts, working on monoclonal-antibody therapeutics. After watching three rounds of lay-offs, Deisher decamped to a Seattle company called Zymogenetics, where she became involved in a cardiovascular-biology group.

Soon after she arrived at the firm in 1995, Deisher isolated what seemed to be pluripotent stem cells from adult cardiac muscle. They differentiated, she says, into cell types including heart muscle, skeletal and smooth muscle, connective tissue, skin, bone and cartilage. "People would come into the lab and they would practically start to drool," Deisher recalls. "It was mind-boggling what these cells became." In March 1998 — 8 months before the first report that human embryonic stem cells had been isolated — the company filed a patent application on the cells, with Deisher listed as first inventor.

It was, and still is, a controversial claim. Kenneth Chien, an expert in studies of heart progenitor cells at the Department of Stem Cell and Regenerative Biology at Harvard University in Cambridge,

"She's kind of the Sarah Palin of stem cells."

STEM CELLS IN COURT

The bid to extend federal funding of human embryonic-stem-cell research has sparked a bitter legal battle.

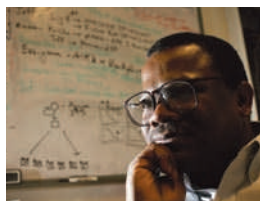
9 MARCH 2009

President Barack Obama rescinds Bush-era restrictions and sets a policy allowing liberalized funding for human embryonic-stem-cell research.



19 AUGUST 2009

Plaintiffs James Sherley, Theresa Deisher, 'embryos' and various others file a lawsuit contesting the legality of funding human embryonic-stem-cell research.



27 OCTOBER 2009

The case is dismissed when the District of Columbia circuit court rules that the plaintiffs have no standing in the case.

25 JUNE 2010

The US Court of Appeals grants standing to Deisher and Sherley alone owing to the competition for limited NIH resources.

23 AUGUST 2010

Judge Royce Lamberth of the District of Columbia Circuit Court grants a preliminary injunction ordering the termination of federal funding to embryonic-stem-cell research under the new regulations.



Massachusetts, says that "nobody has been able to identify a truly pluripotent stem cell from any adult mammalian heart".

Many of her colleagues at Zymogenetics reacted with "ferocious hostility", Deisher says. She recalls one scientist who cornered her, spittle flying from her mouth, shouting: "Adult stem cells do not exist outside the haematopoietic system! Who the blank do you think you are, God?" Deisher was ordered, she says, to stop working on the cells.

The company abandoned the patent application in 2004, but Deisher remains unapologetic about her claims. The website for AVM proclaims: "Dr. Deisher was the first person world-wide to identify and patent stem cells from the adult heart. Her discovery remains one of the most significant discoveries in the area of stem cell research." And the vehemence with which colleagues resisted "made me open my eyes", Deisher says, to the very real — and, she says, unscientific — passions that can infect defenders of scientific orthodoxy. Science, she reasoned, was not so objective after all.

It was a second formative experience for her. Deisher had returned to religion, tentatively, in the early 1990s. Now, her disillusionment with colleagues at Zymogenetics "led me back deeply and profoundly", she says. She left the company for Immunex — which was acquired by Amgen in 2002. Human embryonic stem cells were back in the news, as president George W. Bush defined a policy that allowed federal funding for research on a score of existing cell lines. For Deisher, it was a score too many. "I was extremely disappointed," she says. She felt the policy encouraged an unmerited hype around embryonic cells that deprived adult-stem-cell therapies of support.

Through a friend of her parents, Deisher came into contact with Sharon Quick, a local doctor and conservative activist, who invited her in 2006 to speak on a televised panel about stem-cell research. Murry had also been invited to speak. He recalls Deisher reading prepared remarks about human-embryonic-stem-cell research. "There was a lot of misinformation in there." Her talk, he says, "didn't educate and focus. It obfuscated and frightened."

In response to Murry's criticism, Deisher sent *Nature* a copy of the talk. It argues that human embryonic stem cells could provoke an immune response and form teratomas (tumours containing various types of cell); claims that safe, "clinically proven" alternatives exist; and categorically dismisses any potential promise embryonic cells may offer: "There is no commercial, clinical or research utility in working with human embryonic stem cells." The event put Deisher on the map for anti-embryonic-stem-cell activists. It also led her to a third transformative moment in her advocacy.

In early 2007, Deisher was invited to speak to a group of Republican state lawmakers in Olympia, Washington. One of the other speakers was a mother who had adopted a frozen embryo from a fertility clinic.

The resulting child, a girl then four years old, stood beside her.

Deisher was transfixed. It was, she says, "the turning point to become less scientific about it, and actually feel emotion, and a stronger sense of commitment".

It was this commitment that led Deisher to found AVM in February 2008. The company's mission, in part, is to eliminate the need for embryonic-stem-cell therapies and enable adult-stem-cell companies to succeed by developing, for instance, drugs that promote stem-cell retention in target organs. It is also working on alternatives to vaccines currently produced using cell lines derived from fetuses that had been aborted decades ago. AVM has five members of staff, all of whom are unpaid, and occupies three rooms in a former nurses' dormitory.

She financed AVM with her retirement savings, and with proceeds from the sale of her house. In 2009, an equity offering raised an additional \$225,000 from 'angel' investors. Deisher's non-profit group, the Sound Choice Pharmaceutical Institute, is housed in the same premises and is staffed by four people. Last year, the institute won a \$500,500 two-year grant from the MJ Murdock Charitable Trust, based in Vancouver, Washington, to study whether residual human DNA in the measles, mumps and rubella (MMR) vaccine might trigger autism. Stanley Plotkin, emeritus professor at the Wistar Institute in Philadelphia, Pennsylvania, and inventor of the rubella vaccine, calls the idea "off the wall". "The whole idea, in my view, is just pernicious and just raises a spectre which has been redundantly disproven." John Van Zytveld, a senior fellow at the Murdock trust, who oversees its science grants, says that Deisher's proposal "came back with a strong [peer] review and so we opted to support it".

A CALL TO ARMS

In the spring of 2009, Deisher got a call from Sam Casey, a lawyer then based in Fairfax, Virginia, who was representing Do No Harm, a coalition opposed to human-embryonic-stem-cell research. The US National Institutes of Health (NIH) in Bethesda, Maryland, had just issued draft guidelines proposing to open up funding for the research, complying with an executive order from President Barack Obama. Casey enlisted Deisher to help write the group's response.

He also told her that he was laying plans for a lawsuit if the final guidelines remained substantially unchanged from the draft. The suit would assert that the guidelines contravened an existing law, the Dickey-Wicker amendment, which prohibits federal funding of research in which human embryos are destroyed.

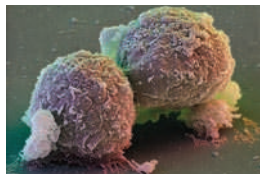
NATURE.COM
For a podcast discussion with the author, see:
go.nature.com/zlm4aq

The NIH published its final guidelines on 6 July 2009, allowing financial support for work on human embryonic stem cells derived ethically from leftover embryos at fertility clinics, but not for work that went into their derivation. "I was

OBAMA: D. MILLS/NEW YORK TIMES/REXUS/EVINE; SHERLEY: C. SUZUKI/AP; LAMBERTH: C. DHARAPAK/AP; CELLS: D. SCHARF/SCIENCE FACTION/CORBIS

9 SEPTEMBER 2010

The US Court of Appeals issues a stay on the injunction allowing federal funds to flow once again to embryonic-stem-cell researchers until the legality of the injunction can be determined.



6 DECEMBER 2010

The Court of Appeals hears oral arguments.

Currently the District of Columbia district court could issue a summary judgement on the legality of the NIH guidelines or the US Court of Appeals could determine the fate of the preliminary injunction. Embryonic-stem-cell research could be halted again. Either way, the case is expected to go before the Supreme Court.

very disappointed,” Deisher says. “I had hoped and thought that they would listen.”

Soon afterwards, Casey, now with the Jubilee Campaign in Washington DC, called Deisher. He told her that the lawsuit was going ahead, and asked her to be one of the plaintiffs. She spent several weeks pondering her decision. “There are huge ramifications to being involved in a lawsuit,” she says. “It is frightening to speak out. I don’t care for the notoriety.” Deisher was also keenly aware that James Sherley had signed on as a plaintiff. She had never met him, but she had followed his widely publicized tenure dispute with MIT. She worried about how a public association with him would affect her reputation.

She made it clear to Casey that if he wanted her as a plaintiff, a high-profile, Sherleyesque approach was out of bounds. “No theatrics, no histrionics, no hunger strikes,” she says. It was agreed, and Deisher joined the suit. Her co-plaintiffs included ‘embryos’; an embryo adoption agency called Nightlife Christian Adoptions; the Christian Medical and Dental Association based in Bristol, Tennessee; and individuals wishing to adopt embryos.

The lawsuit, filed in August 2009, was barely noted by the press. And when, in October that year, District of Columbia District Court Judge Royce Lamberth ruled that none of the plaintiffs had standing to sue, Deisher received the news with a measure of relief. She could return to her preferred focus: her children and her work.

But in June 2010, the US Court of Appeals for the District of Columbia Circuit ruled that Deisher and Sherley alone should be granted standing to sue, because, as adult-stem-cell researchers, they were in danger of ‘imminent’ injury. The court reasoned that by allowing federal funding of embryonic-stem-cell research the NIH increased competition for its limited funds, making it harder for adult-stem-cell researchers to win grants. The appeals court then sent the case back to Lamberth. Deisher was concerned. “It’s a little unnerving to know that you are the only two with standing.”

Unlike Sherley, Deisher has never applied for an NIH grant — as some opponents are quick to point out. She contends that she is still hurt by the guidelines, just as, by her reasoning, all adult-stem-cell researchers are hurt by the NIH’s deliberate focus on embryonic stem cells. Moreover, she says, “I would like to, I intend to and I plan to” apply for NIH grants.

It is hard to argue that adult-stem-cell researchers are at a disadvantage, however. Numbers provided by the NIH show that since 2002, when it first funded a human-embryonic-stem-cell grant, the agency

has spent more than four times as much — \$2.3 billion — on research with non-embryonic human stem cells. Nor has the money for non-embryonic work dwindled as embryonic funding has grown; in 2003, the NIH spent \$191 million on adult-stem-cell research in humans; last year, it spent \$388 million.

Deisher responds that the United States lags in clinically testing new therapeutic uses for adult-stem-cells, instead focusing on well established indications such as leukaemia and lymphoma. Thirty-nine percent of adult-stem-cell trials for ‘unconventional’ indications registered with clinicaltrials.gov take place in the United States, compared with 71% of trials for ‘conventional’ uses. Defenders of the NIH say that lax regulatory and safety hurdles in some countries may explain the discrepancy. Sean Morrison, director of the Center for Stem Cell Biology at the University of Michigan in Ann Arbor, works on adult and embryonic stem cells, and says that “the idea that the NIH is biased against adult-stem-cell research is ridiculous”.

THE HAMMER DROPS

On 23 August 2010, Lamberth issued a preliminary injunction siding with the plaintiffs. That immediately shut down federally funded human embryonic experiments, leaving the research community reeling and angry. Deisher’s phone began ringing off the hook, with queries from reporters around the world. The next day, walking into her office in a building that shares space with other research groups, she was prepared for dirty looks. But “if I got them, I didn’t notice. The response was overwhelmingly positive.”

Deisher made a hastily arranged trip to Washington DC the next week. There, she met Sherley for the first time, during an hours-long strategy session at the offices of Gibson, Dunn and Crutcher, the DC law firm arguing the case. “I asked lots of questions,” Deisher says. (Sherley “is a very nice man”, she adds. “He’s a good scientist.”)

It would be 17 days from the preliminary injunction before a stay from the appeals court allowed embryonic-stem-cell research to resume. Since then, the lawsuit has been proceeding on two tracks. At the lower, district court, Judge Lamberth is considering both sides’ requests for a

speedy, ‘summary’ judgement on whether the NIH’s guidelines are legal. The higher court, the Court of Appeals for the District of Columbia Circuit, which resides one level below the Supreme Court, is considering whether Lamberth met the legal standard for granting the preliminary injunction. Either court could rule at any time, and no matter what the decisions, appeals are expected (see ‘Stem cells

in court’). The case has taken “emotional energy”, Deisher says, but not a great deal of her time. She has not hung on every one of its twists and turns. In many ways, her life goes on unchanged.

Old friends, for example, remain old friends. Two former high-school classmates who recently visited Deisher at her office both adamantly oppose her position on the research, but greet her with evident warmth. “I can say wholeheartedly that I am envious of her passion,” says one. But later, she e-mailed to ask that her name be withheld from this article. “I cannot afford to have a search engine associate me with an individual whose actions are in such opposition to the beliefs of my personal and professional community,” she wrote.

The biggest lesson Deisher has learned from the lawsuit, she says, is “how many scientists are against [human-embryonic-stem-cell research]. I did not know that. I did not expect the level of support and encouragement that I have received.” The extent of that support may be tested if the Court of Appeals for the District of Columbia Circuit, when it rules on the issue, agrees with Deisher. If it does, it will shut down hundreds of human-embryonic-stem-cell experiments once more — possibly for good. ■

Meredith Wadman is a reporter for Nature based in Washington DC.

“It is frightening to speak out. I don’t care for the notoriety.”



Since the turn of the twentieth century, zoologists have set out from coastal marine stations at dawn to sieve peppercorn-sized worms from sea-bottom muck. These creatures, called acoels, often look like unremarkable splashes of paint when seen through a microscope. But they represent a crucial stage in animal evolution — the transition some 560 million years ago from simple anemone-like organisms to the zoo of complex creatures that populate the world today.

There are about 370 species of acoel, which gets its name because it lacks a coelom — the fluid-filled body cavity that holds the internal organs in more-complex animals. Acoels also have just one hole for both eating and excreting, similar to cnidarians — a group of evolutionarily older animals containing jellyfish and sea anemones. But unlike the simpler cnidarians, which have only an inner and outer tissue layer, acoels have a third, middle tissue layer. That is the arrangement found in everything from scorpions to squids to seals, suggesting that acoels represent an intermediate form.

That hypothesis has gained considerable support in recent years, but a report published in *Nature* this week¹ is causing scientists to rethink the storyline. The study by an international team of researchers, who used new analytical techniques and data, removes acoel worms from their position near the trunk of animal evolution and instead places them closer to vertebrates (see ‘Competing views of animal evolution’).

The rearrangement has triggered protests from evolutionary biologists, who are alarmed that they may lose their key example of that

An obscure group of tiny creatures takes centre stage in a battle to work out the tree of life.

BY AMY MAXMEN

crucial intermediate stage of animal evolution. Some researchers complain that the evidence is not strong enough to warrant such a dramatic rearrangement of the evolutionary tree, and claim that the report leaves out key data. In any case, the vehemence of the debate shows just how important these worms have become in evolutionary biology.

“I will say, diplomatically, this is the most politically fraught paper I’ve ever written,” says Max Telford, a zoologist at University College London and last author on the paper.

The debate focuses on where acoels fit in the family tree of bilaterians, three-layered animals with bilateral symmetry. Biologists divide these animals into two branches. The larger group,

called protostomes, contains invertebrates such as earthworms, squids, snails and insects. The smaller group, known as deuterostomes, includes both vertebrates and invertebrates, such as sea urchins, humans and fish.

Zoologists have generally placed acoels on the earliest branch of the bilaterians — before the split between protostomes and deuterostomes — because the worms lack so many key features such as a

separate mouth and anus, a central nervous system and organs to filter waste. Although the position of acoels has moved around a bit over the decades, a DNA analysis in 1999 (ref. 2) and several since then have placed them back in their earlier spot. In particular, a genetic study of 94 organisms in 2009 solidified the conclusion that acoels belonged at the very base of the bilaterians³. That study, led by Andreas Hejnol, a developmental biologist at the Sars International Centre for Marine Molecular Biology in Bergen, Norway, confirmed that acoels and their kin occupied an intermediate spot between cnidarians and the more-complex bilaterians.

“I suddenly had the feeling that everything had finally fallen into place,” says Claus Nielsen, an evolutionary biologist at the Natural History Museum of Denmark, who has followed acoels for 40 years as they wandered across the tree of life.

SHAKING THE TREE

But the study by Telford and his colleagues¹ has shaken the tree again and placed acoels within the deuterostome branches, next to the echinoderms (which include sea urchins) and acorn worms. Their genetic analyses suggest that the acoels — and a marine worm named *Xenoturbella* — descended from a more complex ancestor and lost many of the features seen in other deuterostomes.

The researchers used several approaches and examined three independent data sets to come to their conclusions. First, they reanalysed data from Hejnol’s 2009 study³, using 66 species instead of 94. Hervé Philippe, a bioinformatician at the University of Montreal in Quebec, Canada, and first author of the *Nature* paper¹, says that the team removed species that had

“THIS IS THE MOST POLITICALLY FRAUGHT PAPER I’VE EVER WRITTEN.”

NATURE.COM

Read Telford and colleagues’ study at: go.nature.com/hrwepf

A. WALLBERG

incomplete genetic data or were ‘fast-evolving’ — meaning that some of their genes had accumulated many changes, when compared with genes from animal groups that emerged around the same time. Phylogenetic computer programs have a well-known problem with these kinds of species and tend to group them together even though they are not related.

Philippe and his co-workers used a more sophisticated mathematical model to analyse sequence evolution, which helped to minimize this problem. Without this model and careful species selection, Philippe says, acoels can fall at the base of the animal tree.

After analysing sequences from nuclear DNA, the group made a separate evolutionary tree based on genes in mitochondria. They also studied microRNAs, which regulate gene expression but do not code for proteins. According to co-author Kevin Peterson, a palaeontologist at Dartmouth College in Hanover, New Hampshire, microRNAs are particularly useful for studying deep evolutionary relationships. The team found that acoels have a type of microRNA known to be specific to deuterostomes, suggesting that they are related.

The authors acknowledge that no single data set clinches the case for placing acoels within

the deuterostomes. But taken together, says Telford, “the fact that our evidence points in the same direction makes me think it’s right”.

If acoels do fit within the deuterostomes, the worms must have evolved from an ancestor with a central nervous system, a body cavity and a through-going gut that connected an anus and mouth — features seen in existing deuterostomes. So researchers would need to explain how acoels and *Xenoturbella* lost those and other characteristics. They would also be left to search for another primitive-looking lineage that represents the evolutionary step between jellyfish-like animals and bilaterians. (If one even exists. Peterson says that many complex features may have emerged all at once.)

Some researchers are not ready to give up on the old ideas of where acoels fit. “I’m sad about their paper, but I’m not upset,” says Hejnol. “I’d be upset if their analysis was excellent and it meant we lost a representative animal to bridge an important transition in the tree of life.”

Hejnol and his colleagues have doubts about the reliability of the tree that Telford and his team built from nuclear genes, which

is their main evidence. Critics say that the key branches of the tree are not as statistically strong as they should be.

Because of this, Brian O’Meara, a phylogeneticist at the University of Tennessee in Knoxville, calls the new tree “suggestive, but not definitive”.

The study has also come under fire for leaving out data that some scientists say would

have weakened the researchers’ conclusions. An author on the paper had previously analysed a species of worm closely related to acoels known as *Meara stichopi*, and did not find deuterostome microRNA. But the authors

defend their decision to keep *M. stichopi* out of their microRNA analysis owing to concerns about the quality of those data.

Moreover, not everyone is convinced by the power of microRNA analysis, which has only recently been adopted for evolutionary studies. This report marks the method’s most high-profile appearance yet as a tool to resolve relationships. Because microRNAs can be lost during evolution, it is possible that the deuterostome microRNA in acoels originated in the ancestor of all bilateral animals but was later lost in the protostome line.

With so much at stake, researchers are keen to resolve the issue. The US National Science Foundation has been specifically soliciting proposals that target deep divergences in evolutionary history, as part of an initiative called Assembling the Tree of Life, says Tim Collins, a programme director at the foundation. “We’ve done a good job within groups, but we’ve had a hard time reconstructing the deepest branches of the tree of life,” he says. “These are the events that happened in a relatively short time compared with the amount of time that has passed since then, which makes things hard.”

Last summer in Kristineberg, Sweden, Hejnol and Telford shared a room while teaching a class together. They debated their differences and discussed an ongoing joint project that might settle them: sequencing the full genomes of an acoel, a species of *Xenoturbella* and the controversial *M. stichopi*. With that influx of new genomic information, the researchers are confident that they can reach an agreement about where acoels fit in evolutionary history.

“We’re talking about a very close result with a humongous impact,” says Hejnol, of the newly proposed tree. “The good thing is, we know how to resolve this issue.” ■

Amy Maxmen is a freelance writer based in New York City.

1. Philippe, H. *et al.* *Nature* **470**, 255–258 (2011).
2. Ruiz-Trillo, I., Riutort, M., Littlewood, D. T., Herniou, E. A. & Baguña, J. *Science* **283**, 1919–1923 (1999).
3. Hejnol, A. *et al.* *Proc. R. Soc. B* **276**, 4261–4270 (2009).

“I’M SAD ABOUT THEIR PAPER, BUT I’M NOT UPSET.”

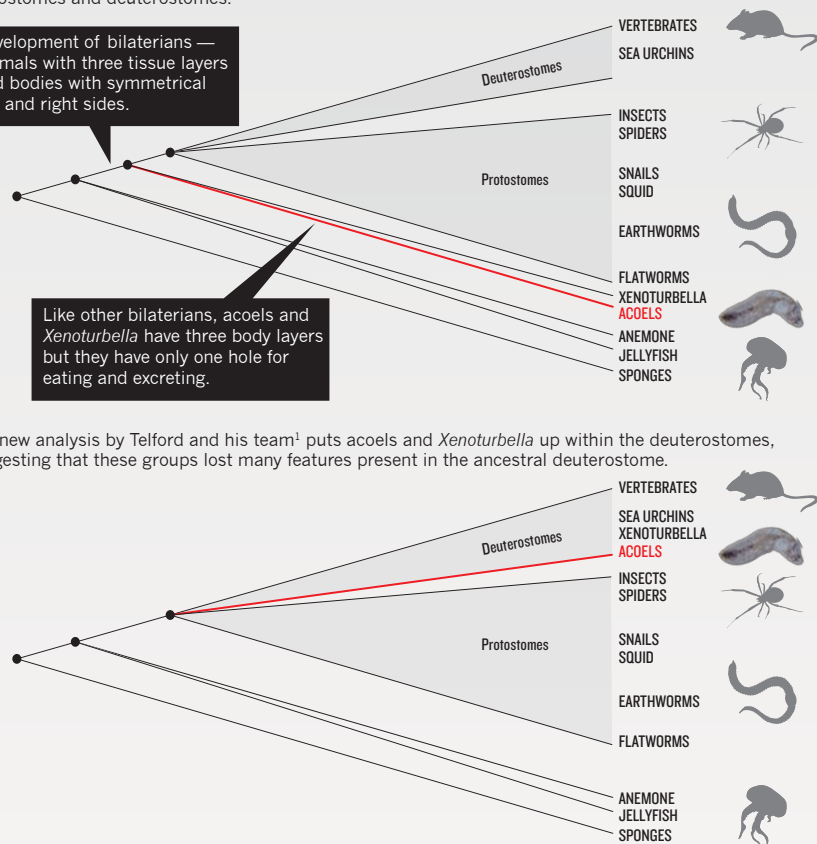
COMPETING VIEWS OF ANIMAL EVOLUTION

The traditional view of acoels places them at the base of the bilaterians, before the evolution of animals with a separate mouth and anus. After acoels and *Xenoturbella* split off, bilaterians diverged into protostomes and deuterostomes.

Development of bilaterians — animals with three tissue layers and bodies with symmetrical left and right sides.

Like other bilaterians, acoels and *Xenoturbella* have three body layers but they have only one hole for eating and excreting.

The new analysis by Telford and his team¹ puts acoels and *Xenoturbella* up within the deuterostomes, suggesting that these groups lost many features present in the ancestral deuterostome.



COMMENT

ANTHROPOLOGY Call for unity in the science of human beings **p.166**



GENETICS Reviewed: two primers on personal genomics **p.169**

POLICY Sanitation, not vaccination, is most important in Haiti **p.175**

OBITUARY Jack Oliver, key player in proof of plate tectonics, remembered **p.176**

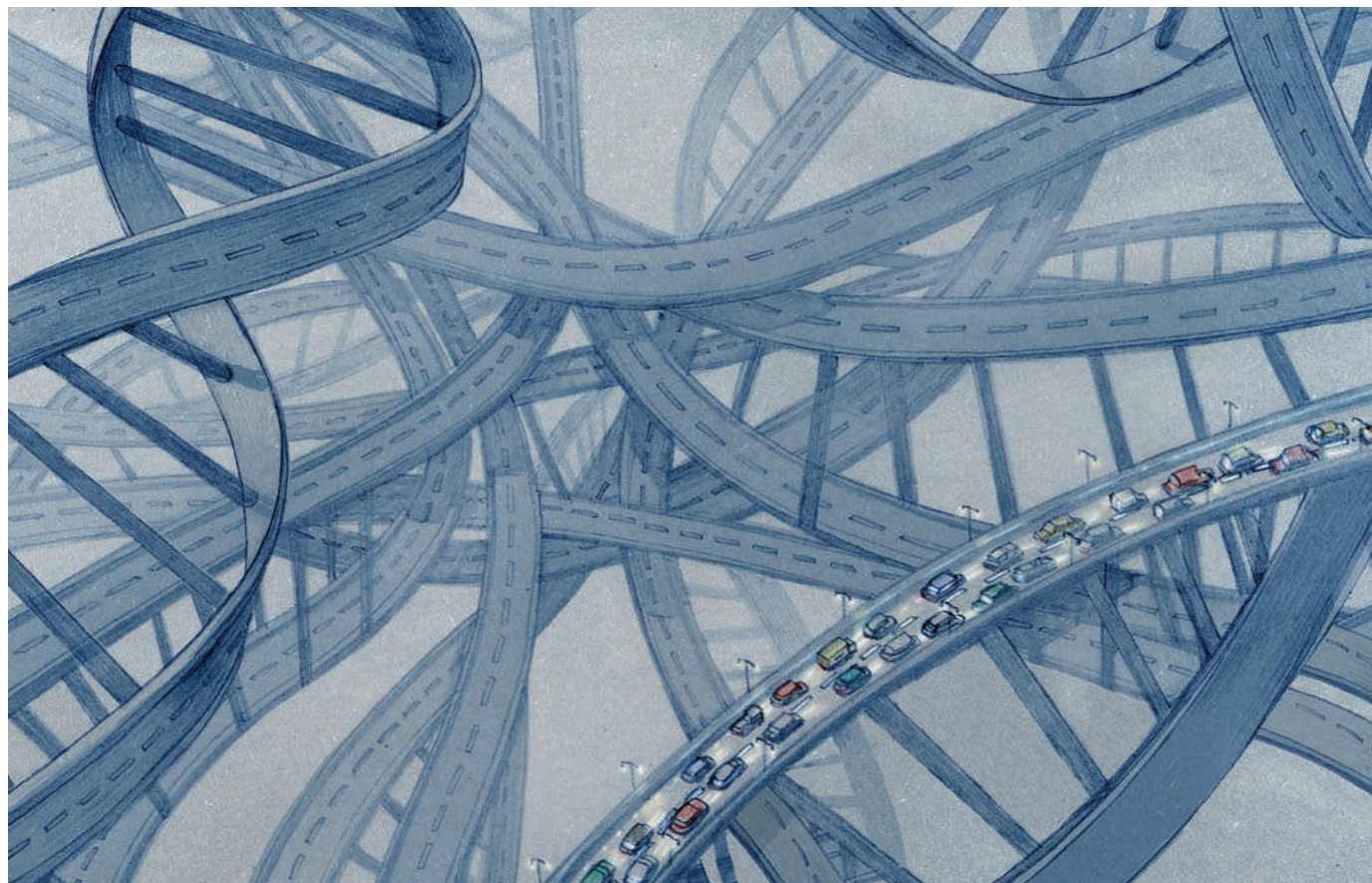


ILLUSTRATION BY JONATHAN BURTON

Too many roads not taken

Most protein research focuses on those known before the human genome was mapped. Work on the slew discovered since, urge **Aled M. Edwards** and his colleagues.

When a draft of the human genome was announced in 2000, funders, governments, industry and researchers made grand promises about how genome-based discoveries would revolutionize science. They promised that it would transform our understanding of human biology and disease, and provide new targets for drug discovery. Yet more than 75% of protein research still focuses on the 10% of proteins that were known before the genome was mapped — even though many more have been genetically linked to disease.

We performed a bibliometric analysis to assess how research activity has altered over time for three protein families that are central in disease and drug discovery: kinases, ion channels and nuclear receptors. For all three, we found very little change in the pattern of research activity — which proteins are associated with the highest number of publications — over the past 20 years¹.

NATURE.COM
Protein mapping gains a human focus:
go.nature.com/vbqctf

Even those proteins that have been directly associated with disease

remain 'hidden in plain sight', with scientists proving very reluctant to study them.

Where there has been a shift in research activity, it was often spurred by the emergence of tools to study a particular protein, not by a change in the protein's perceived importance. We believe that ensuring high-quality tools are developed for all the proteins discovered may be all that is needed to drive research into the unstudied parts of the human genome — even within funding and peer-review systems that are inherently conservative.

We searched for mention of every human ▶

► kinase, ion channel and nuclear receptor in either the title, abstract, keywords or 'MeSH' terms (used for indexing articles in Medline and PubMed) in the almost 20 million papers published between 1950 and 2009. We discovered that for all three classes of protein, the same small fraction of family members have remained 'the favourites' for nearly 20 years (see 'Fondling our problems').

For instance, the human genome encodes more than 500 protein kinases, of which hundreds have been shown to have genetic links with human diseases. Yet around 65% of the 20,000 kinase papers published in 2009 focused on the 50 proteins that were the 'hottest' in the early 1990s. Similarly, 75% of the research activity on nuclear hormone receptors in 2009 focused on the 6 receptors — out of the 48 encoded in the genome — that were most studied in the mid 1990s (ref. 1).

BIASED APPROACH

Although academics may be surprised by the magnitude of this research bias, they generally acknowledge its existence. It was first identified in kinase research² in 2008 and last year its effects were demonstrated in kinase drug discovery³. But a common assumption is that previous research efforts have preferentially identified the most important proteins. The evidence doesn't support this.

Patterns of gene expression and links between DNA sequences and breast cancer suggest, for instance, that 11 protein kinases are key nodes in the signalling pathways underlying the disease. Yet in the 2009 literature, one of these kinases, CDC2, received more attention than seven of the others combined, and three received just one mention. Likewise, various genetic approaches, including genome-wide association studies, have directly linked 37 of the 48 human nuclear receptor genes to disease. Among these, more than half of the total research

activity in 2009 was focused on just three. These three were also 'top of the nuclear receptor charts' in the 1990s.

Why the reluctance to work on the unknown? As the Nobel-prizewinning biochemist Roger Kornberg put it, scientists are wont to "fondle their problems": they have a natural tendency to dig deeper into their areas of expertise. Plus, funding and peer-review systems are risk-averse; funders and reviewers alike are less willing to support research on unstudied proteins, for which it is often harder to explain the rationale and significance. Moreover, the time frames associated with academic promotion and training encourage researchers to focus on systems that are likely to generate results rapidly, and for which research infrastructure and methods are already available.

Some funders are developing strategies to address the conservative nature of peer review. The Wellcome Trust, the largest non-governmental funder of biomedical research in the United Kingdom, for instance, is withdrawing its project grants in favour of providing longer-term support to outstanding investigators. And many universities are examining the pitfalls of their current reward systems. Unfortunately, institutional systems are ponderously slow to change. So what else can be done?

To establish a protein's function, and especially the details of how it works and its suitability for drug discovery, molecular biologists draw on an arsenal of tools. For instance, antibodies can help them identify where in the body the protein is being expressed; chemical inhibitors can be used to block a protein's activity in human cells and in animal models. These antibodies and small molecules also provide a launch pad for the development of new medicines by the biotechnology and pharmaceutical industries. Yet because of the cost and time

required to generate and characterize such tools, they are currently available for only a handful of well-studied proteins.

Our analysis of publication patterns for the human nuclear hormone receptor family suggests that making such tools readily available for all proteins could dramatically shift the balance in biomedical research.

WAKE-UP CALL

Nuclear receptors are transcription factors that bind small signalling molecules, such as steroids and hormones. Genetic data now suggest that all the receptors are directly or indirectly linked to human disease. About 30 of the family members were discovered at around the same time in the 1990s, allowing us to compare publication trends for numerous related proteins over time. We know exactly when the receptors were cloned, when genetic links with diseases were established and when research tools (in this case, chemical probes)^{4,5} became available.

When the 'novel' nuclear receptors were identified in the 1990s, all the family members were thought to have therapeutic potential.

"Making protein-based research tools readily available must be a major objective in the decade to come."

Interest developed most rapidly in those that were found to have genetic links to disease⁶⁻⁸ or that had interesting knockout phenotypes⁹, such as infertility. However, over the next 15 years, research activity refocused on a subset of 8 of these receptors.

From a genomics point of view, these 8 are no more interesting than any of the other 29 with known links to disease.

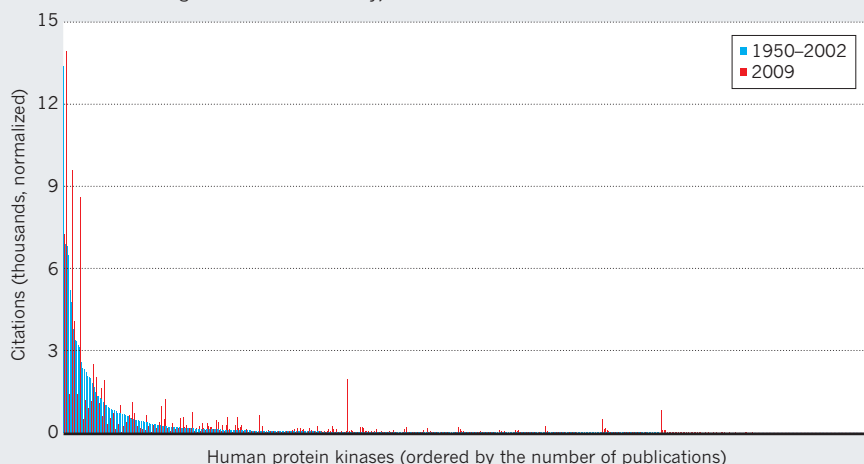
To our knowledge, the only connection among these 8 receptors is that for each there is a widely available, high-quality chemical probe that either enhances the receptor's activity or dampens it. In short, where high-quality tools are available (often commercially), there is research activity; where there are no tools, there is none (see 'Tools are telling').

Several other observations are consistent with the ideas that the availability of chemical probes for a given receptor dictates the level of research interest in it, and that the development of these tools is not driven by the importance of the protein. For instance, large and sustained increases in the rate of publications mentioning a nuclear receptor usually followed, not preceded, the release of a chemical probe.

Our findings should serve as a wake-up call to the biomedical and pharmaceutical research communities. Granting systems must be more daring, institutions must foster and reward risk, and the entire biomedical community must play down the legacy of the literature and let new evidence guide

FONDLING OUR PROBLEMS

Researchers' 'favourite kinases' have remained the same for decades with a few exceptions (kinases linked to diseases of great interest to industry).



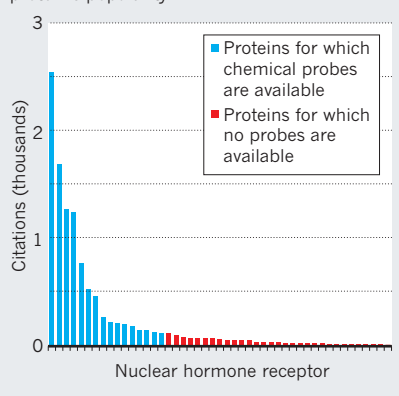
research. Genome-wide tools such as the DNA microarrays used in association studies have allowed geneticists to ignore preconceived ideas about disease mechanisms and pursue a remarkably successful broad-brush approach; this approach should be embraced more generally.

Our data also indicate that high-quality, readily available research tools can dramatically facilitate exploratory biomedical research. Funders such as the Wellcome Trust and the US National Institutes of Health have allocated some funding to tool-generating projects, but perhaps not enough. Part of the problem is that, unlike the high-energy physics community, which endorses the creation of large resources, the biomedical community often views projects focused on tool creation with some disdain, for lacking the elegance of 'real' science.

The budgets required also incite a visceral reaction. For example, the level of funding needed to develop even one chemical probe is enormous. Although it is only a fraction of the US\$100 billion spent on biomedical research each year — about several million dollars — it is huge compared with the amount customarily allocated to an individual scientist. Finally, the risk is significant. Large-scale efforts are not guaranteed to succeed; they require expertise in science

TOOLS ARE TELLING

The availability of research tools influences a protein's popularity.



and management, as well as collaboration between disciplines, between public and private sectors and — to avoid duplication of effort — even between countries.

Much of the work that has emerged from exploring the human genome over the past ten years lies fallow. Challenges notwithstanding, making protein-based research tools readily available must be a major objective in the decade to come. ■

Aled M. Edwards is Banbury professor at the

University of Toronto, Toronto, Ontario M5G 1L7, Canada. **Ruth Isserlin** and **Gary D. Bader** are at the Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. **Stephen V. Frye** is at the Center for Integrative Chemical Biology and Drug Discovery, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. **Timothy M. Willson** is at GlaxoSmithKline, Research Triangle Park, North Carolina 27709, USA. **Frank H. Yu** is in the Program in Neurobiology and Dental Research Institute, School of Dentistry, Seoul National University, Seoul, 110-749, South Korea.

e-mail: aled.edwards@utoronto.ca

The authors declare competing financial interests: details accompany this article online at go.nature.com/kgw3oz.

1. Isserlin, R. *et al.* Preprint at <http://arxiv.org/abs/1102.0448> (2011).
2. Grueneberg, D. A. *et al.* *Proc. Natl Acad. Sci. USA* **105**, 16472–16477 (2008).
3. Fedorov, O., Müller, S. & Knapp, S. *Nature Chem. Biol.* **6**, 166–169 (2010).
4. Frye, S. V. *Nature Chem. Biol.* **6**, 159–161 (2010).
5. Willson, T. M. & Moore, J. T. *Mol. Endocrinol.* **16**, 1135–1144 (2002).
6. Muscatelli, F. *et al.* *Nature* **372**, 672–676 (1994).
7. Yamagata, K. *et al.* *Nature* **384**, 458–460 (1996).
8. Achermann, J. C., Ito, M., Ito, M., Hindmarsh, P. C. & Jameson, J. L. *Nature Genet.* **22**, 125–126 (1999).
9. Lee, S. L. *et al.* *Science* **273**, 1219–1221 (1996).



Anthropologists unite!

Anthropology isn't in the crisis that parts of the media would have you believe, but it must do better, argue **Adam Kuper** and **Jonathan Marks**.

In December 2010, *The New York Times* reported^{1,2} that the term ‘science’ had been dropped in a new long-range plan of the American Anthropological Association (AAA). Where once the association had dedicated itself “to advance anthropology as the science that studies humankind in all its aspects”, it now promised rather “to advance public understanding of humankind in all its aspects”.

The reporter suggested that this brought to a head an epic struggle in the discipline between the true scientists and their foes. Frank Marlowe, president-elect of the Evolutionary Anthropology Society (a branch of the AAA that is curiously independent of its long-standing Biological Anthropology Section), was quoted as saying: "we evolutionary anthropologists are outnumbered by the new cultural or social anthropologists, many but not all of whom are postmodern,

which seems to translate into antiscience.”

The new long-range plan also provoked rumblings of discontent (still ongoing) in the blogosphere, and the association's executive committee scrambled somewhat belatedly to reassure the public — and its own members — that it had all been a misunderstanding. They had not intended to cast doubt on the scientific character of the discipline. And in fact the same committee had come up with a simultaneous text entitled 'What is Anthropology?', which describes anthropology unambiguously as a science.

Apparently, a committee had floundered in trying to come up with an agenda for anthropology that was baggy enough to accommodate its very various research programmes. Is this news? Indeed it is, but not, as the bloggers and *The New York Times* suggest, because an anti-science conspiracy has hijacked American anthropology. The real shocker is that

anthropologists cannot agree on what the discipline is about. Many, probably most, anthropologists have walked away from their traditional mission, which is to build a truly comparative science of human variation. We need to work out where we are now heading.

ROOTS AND BRANCHES

The reason that the AAA got into such a pickle is that — like geography, even perhaps like biology — anthropology is a nineteenth-century discipline that fragmented, spawning a variety of specializations. Biological anthropology, archaeology and the various traditions of ethnography are bundled together in many university departments and professional associations such as the AAA and, in Britain, the Royal Anthropological Institute. However, relationships are often distant. The biologists do genetics, or neuroscience, or primatology, or chase up

new developments in evolutionary theory. They show little interest in archaeology — except perhaps the archaeology of very ancient humans — or in ethnography, except for snippets of information about sex and violence. Some do seem to feel that if only they could spare the time they would be able to knock some evolutionist sense into cultural anthropology. But they are too busy.

Meanwhile, the ethnographers agree that their first task is to document the great diversity of human ways of life. Generalizations about human nature should not be based on a single report of Amazonian violence, or Tibetan polyandry, or woman–woman marriage among the Lovedu of South Africa. But they do not agree on how to make sense of the customs of faraway peoples. Social anthropologists engage with models and theories current in the social sciences (ideally, although they seldom keep up as well as they should). Some cultural anthropologists aim rather to understand and translate, and they look for inspiration to literary theorists and philosophers (preferably French, even if they have to be read in often impenetrable translations).

For a long time the main branches of anthropology largely ignored one another, but in the 1980s two radical movements provoked a confrontation. Sociobiologists claimed that genetics was about to revolutionize the human sciences. These would become at last a branch of biology, although the great biologist Ernst Mayr did warn that “the profound differences in social behaviour among human groups, some of them closely related, show how much of this behaviour is cultural rather than genetic”. Sociobiologists also drew on ethology, an older movement that made much of parallels between human and primate — or even insect — behaviour, provoking Sherwood Washburn, a leading biological anthropologist, to comment that human ethology “might be defined as the science that pretends humans cannot speak”.

Inspired by the elegant essays of Clifford Geertz, another new movement appeared centre-stage in the 1980s (in fact another very old movement, in modern dress). Cultural theorists, identifying themselves with the humanities, insisted that foreign ways of thought are resistant to translation, that variation and change characterize even the most isolated populations, and that it is therefore not easy to say what the Bushmen do, or the Trobrianders, or for that matter the English (all of them? Always?), so comparisons are problematic. Some disciples of

Geertz followed that road down to a relativist dead end. All generalizations about human beings were suspect, except for the iron law that culture trumps biology³.

The controversies of the 1980s, which lingered on into the 1990s, often hinged on claims about race, sex and violence, and so they caught the attention of a wider public. In a popular book published in 1928, Margaret Mead had reported that Samoan girls enjoyed sexual freedom, and so experienced an untroubled passage through adolescence⁴. More than half a century later (and after Mead's death), Derek Freeman trashed her account, insisting that the girls were remarkably chaste⁵. Rather mysteriously, the sex life of Samoan girls became a popular test-case for the nature–nurture argument. (Recent commentaries are kinder to Mead than to Freeman, although it has become obvious that neither Freeman nor Mead can be relied on uncritically for the description of Samoan adolescence, let alone for the explanation⁶.)

Young women might find happiness in a liberated sex life, but were young men given rather to violence? Napoleon Chagnon claimed that among the Yanomami of the Amazon, the most violent men got the girls. (And he suggested that, in something like a state of nature, all men are Yanomami under the skin⁷.) His account of these people was challenged by other ethnographers, who reported significant local variation even among the 22,500-strong Yanomami, not least in rates of homicide and the abduction of women⁸. In any case, the Yanomami are not typical even of the most isolated, small-scale, technologically limited societies. Many ethnographies document easy-going gender relationships between hunter-gatherers, from Alaska to the Kalahari Desert, or offer historical accounts of peace-loving Indian chiefs with many wives, presiding over a monastic soldiery.

Race was altogether a more serious matter, but on this the anthropologists were not fundamentally divided. The 1994 publication by psychologist Richard J. Herrnstein and political scientist Charles Murray of their book *The Bell Curve*⁹ provoked a national debate about race and inequality. The AAA and the American Association of Physical Anthropologists issued parallel statements summarizing the scientific understanding on race. In brief, they agreed that human variation is structured bioculturally, clinally and locally. Nothing corresponding to the zoological subspecies exists within extant *Homo sapiens*. Individuals and groups of people do indeed differ biologically. However, social inequalities are overwhelmingly the product of political and economic history, not of microevolution.

In the course of the feuding 1980s, several flagship anthropology departments in the United States split up. The biologists joined

faculties of science or medicine. Cultural anthropologists allied themselves with the humanities. Archaeologists sought shelter where they could. In Europe the main branches of anthropology had gone their own ways after the Second World War. It now seemed as though the Americans were belatedly following the same route. However, in the new millennium, the brief and localized trend reversed itself. This is because there is a student demand for the whole package, the study of human origins, history and diversity.

Today, anthropologists may teach more or less happily in interdisciplinary teams, but they seldom collaborate in research projects that breach their disciplinary specialties.

“There is a need for a truly comparative science of human beings throughout their history, and all over the world.”

In the past few years they have drifted to a sadder-but-wiser default position, some documenting the range of differences in human biology, others studying the world of social institutions and belief systems. Only a handful still try to understand the origins and possible connections

between biological, social and cultural forms, or to debate the relative significance of history and microevolution in specific, well-documented instances.

This is a great pity, and not only because the silence of the anthropologists has left the field to blockbusting books by amateurs that are long on speculation and short on reliable information. Anthropologists hardly bother any longer to take issue with even the most outlandish generalizations about human nature. Not their business.

BETTER TOGETHER

To be sure, it is not easy to make general statements about human nature, or even to define it. One obstacle is the often-taken-for-granted opposition between the notoriously — perhaps necessarily — unstable ideas of ‘nature’ and ‘culture’. The human species has been co-evolving with technology for millions of years. Advances in contraceptive techniques have transformed our sexual behaviour. The most fundamentally hard-wired human adaptations — walking and talking — are actively learned by every person, in each generation. So whatever human nature may be, it clearly takes a variety of local forms, and is in constant flux.

The obvious conclusion is that interdisciplinary research is imperative. Yet too few biological anthropologists attend to social or cultural or historical factors. A minority of cultural anthropologists and archaeologists do apply evolutionary theory, or cognitive science, or adopt an ecological

“The real shocker is that anthropologists cannot agree what the discipline is about.”

perspective on cultural variation, or play about with the theory of games, but they feel that they are isolated, even marginalized. And they do not feature in the front line of current debates about cognition, altruism or, for that matter, economic behaviour or environmental degradation, even though these debates typically proceed on the basis of very limited reliable information about human variation. A rare exception is the field of medical anthropology, where cultural anthropologists engage regularly with biologists in studies of HIV and AIDS, or post-traumatic stress disorders, or investigations of folk medical beliefs and practices.

Yet even allowing for their current head-down posture, anthropologists do share a great common cause. They would agree that anyone who makes claims about human nature must learn a lot of ethnography. This does not mean parachuting into the jungle somewhere to do a few psychological experiments with the help of bemused local interpreters, or garnishing generalizations with a few worn and disputed snippets about exotic customs and practices. Unfortunately, very nearly

NATURE.COM
For more on working
in anthropology see:
go.nature.com/3n11uo

all research funding in the human sciences is directed to the study of the inhabitants of



North America and the European Union. Ninety-six per cent of the subjects of studies reported in the leading American psychology journals are drawn from Western industrial societies¹⁰. These represent a minuscule and distinctly non-random sample of humanity.

So there is a need for a truly comparative science of human beings throughout their history, and all over the world. This requires more interdisciplinary team research in anthropology. A good start would be for

anthropologists to read each other's papers, to attend each other's conferences and to debate concrete cases and specific hypotheses. But there is no future in a return to the feuding parties of the 1980s. ■

Adam Kuper is a visiting professor at the Department of Anthropology, London School of Economics, London WC2A 2AE, UK.

Jonathan Marks is in the Department of Anthropology, University of North Carolina-Charlotte, Charlotte, North Carolina 28223, USA.

e-mail: adam.kuper@gmail.com

1. Wade, N. Anthropology a science? Statement deepens a rift. *New York Times* (9 December 2010).
2. Wade, N. Anthropology group tries to soothe tempers after dropping the word 'science'. *New York Times* (13 December 2010).
3. Kuper, A. *Culture: The Anthropologists' Account* (Harvard Univ. Press, 1999).
4. Mead, M. *Coming of Age in Samoa* (William Morrow, 1928).
5. Freeman, D. *Margaret Mead and Samoa* (Harvard Univ. Press, 1983).
6. Shankman, P. *The Trashing of Margaret Mead: Anatomy of an Anthropological Controversy* (Univ. Wisconsin Press, 2009).
7. Chagnon, N. A. *Science* **239**, 985–992 (1988).
8. Albert, B. *Curr. Anthropol.* **30**, 637–640 (1989).
9. Herrnstein, R. J. & Murray, C. *The Bell Curve: Intelligence and Class Structure in American Life* (Free Press, 1994).
10. Arnett, J. J. *Am. Psychol.* **63**, 602–614 (2008).



GENOMICS

Sequence sharing

Peter Border asks how we can protect our personal genomic data while making them available for research.

The Human Genome Project cost US\$3 billion and took 13 years. Today, sequencing machines can churn through a whole human genome in days for a few thousand dollars, making personal genomics increasingly affordable. Two books, Kevin Davies's *The \$1,000 Genome* and Misha Angrist's *Here is a Human Being*, chart the growth of personal genomics and examine its implications.

Davies, a biomedical journalist and founding editor of *Nature Genetics*, focuses on the scientific advances that are enabling more of us to map our genes. Angrist, a geneticist at Duke University, North Carolina, examines the personal and political consequences. Both agree that more scientific work is needed to improve the usefulness of genome data for health care, and that this will require sharing data with researchers. Neither says much about how this might be achieved in practice,

The \$1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine

KEVIN DAVIES

Free Press: 2010. 352 pp. \$26

Here is a Human Being: At the Dawn of Personal Genomics

MISHA ANGRIST

Harper: 2010. 352 pp. \$26.99, £17.99

or how people's concerns about genomic privacy and security can be overcome.

Davies focuses on the developments that replaced automated versions of Sanger sequencing with second-generation, high-throughput machines. He explains the science that underpins the massively parallel, miniaturized approaches used and gives a glimpse of future developments. He acquaints us with DNA microscopes, nanopores and ion sensors, techniques that are vying to form the basis of third-generation sequencers

that might deliver whole genomes in a few minutes for a few hundred dollars.

Angrist is a part of the Personal Genome Project, an ambitious plan to sequence human genomes and make them freely available on the Internet. The project has two underlying principles. First, genome information is most useful when linked to other, phenotypic information about an individual's medical records or family history. Second, it is misleading and unrealistic to guarantee participants' genomic privacy. As Angrist points out, DNA is the ultimate identifier, particularly when it is linked to detailed personal information.

Attempts to sanitize genome data are fraught with difficulty. Angrist discusses James Watson's genome, one of the first to be made public. Watson stipulated he did not want to know if he was at an increased risk of Alzheimer's disease, so asked for the sequence data for the relevant gene, *ApoE4*, to be removed before publication. But this proved futile; researchers showed it is easy to deduce *ApoE4* status by studying the sequence on either side of that gene.

Participants in the Personal Genome Project were given the opportunity to leave out some of their genome data, but most did not. They took part with the expectation that their genomes would become public currency and were warned of the implications. For instance, genetic information might be used to infer paternity, to adversely affect employment or insurance situations, or even to make synthetic DNA that could be planted at a crime scene. Angrist describes how he agonized over going public with his own genetic and medical record, including a family history of breast cancer that might raise future concerns for his two daughters.

Each book discusses the growth of 'spit-kit' companies that offer genome analysis through direct-to-consumer testing. For a few hundred dollars, companies such as 23andMe, Navigenics and Pathway Genomics will analyse a sample of your DNA and provide a report. They also allow you to compare your genome against a maintained online database of sequences and associated traits, from the trivial to the potentially life-changing.

The debate continues as to whether consumer genetic tests should be more closely regulated, with questions about the veracity of the tests and how they should be offered: are the links between the genetic markers and the associated traits robust enough to allow reliable probabilistic assessments? Can these lead to useful health interventions or lifestyle changes? Are consumers or their physicians sufficiently informed to interpret genome information in a meaningful way? Neither

book draws hard and fast conclusions.

Everyone agrees that the coming deluge of genome data will need to

NATURE.COM

The Human Genome at Ten:
go.nature.com/ugle41

be linked to information about an individual's health, environment and family background. However, the use of genome data in research raises privacy and confidentiality issues. The Personal Genome Project model that Angrist describes represents one end of a spectrum. Its participants are advocates of biomedical research, actively involved in the field and able to understand the possible consequences of having their data posted online.

Direct-to-consumer testing is nearer the other end of the spectrum. Each consumer's genome information sits on a secure server maintained by the testing company, and it is up to each person how they share it. The popularity of this approach suggests that people trust the companies to hold their information and like being in control of who can see it. The downside is that the information is less accessible to researchers.

The challenge will be to reconcile people's concerns about genomic privacy and security with the need to allow researchers and clinicians data access. If personal genomics is offered largely by consumer testing, the onus will be on the companies to engage their customers in research projects that improve understanding of how our genomes influence health. Some have started to do this. For instance, 23andMe — the research arm of 23andMe — recruits customers into research projects on pharmacogenomics, Parkinson's disease, sarcoma and aspects of pregnancy.

Alternatively, national or regional health-care systems concerned with disease prevention may offer genomic tests on a more systematic basis. A possible model might be the UK Biobank, which has recruited more than 500,000 people in a study to see how health is affected by lifestyle, environment and genes. Participants divulge medical and lifestyle information and donate biological samples from which researchers can generate gene sequences. The success of the project to date is largely a result of creating a robust, independent framework for ethics and governance, recruiting participants through trusted intermediaries (their physicians) and sharing the data with researchers in a form that hides the identity of individuals.

No amount of data will be useful if you can't interpret what they mean. We are reaching the point at which the cost of interpreting genome information will exceed the cost of generating it, so the challenge ahead will be to make more sense of the data we already have. We will also have to answer the question of whether genome data are personal, in that they are paid for and controlled by individuals, or whether such data are medical, being funded by and accessible to health-care systems. ■

Peter Border is biology and health adviser at the Parliamentary Office of Science and Technology, London, UK.
e-mail: peter.border@googlemail.com

CHEMISTRY

A cultural history of the elements

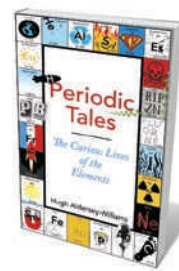
Andrew Robinson enjoys a chemical romp from aluminium to zinc.

The familiar metallic taste of blood was explained scientifically only in the mid-eighteenth century. An Italian chemist and physician in Bologna, Vincenzo Menghini, roasted the blood of various birds, fish and mammals, including humans. He powdered the residue and passed a naturally magnetized lodestone close to the dried blood particles. Some were magnetically attracted, suggesting the presence of iron.

Science writer Hugh Aldersey-Williams successfully repeated Menghini's experiment in his kitchen using blood from chicken livers, a low oven and a moderately strong magnet. But why, he asks, did Menghini imagine that iron would be present in blood? The physician must have known that people with blood disorders were sometimes advised to take iron salts. In the sixteenth century, a principal iron ore was named haematite — the prefix 'haem' being derived from the Greek for blood. Western alchemists paired iron with the red planet Mars. The original connection between iron and blood seems to date to the Romans,



Prince Louis Napoleon's 1856 aluminium rattle.



Periodic Tales: The Curious Lives of the Elements/A Cultural History of the Elements, from Arsenic to Zinc
HUGH ALDERSEY-WILLIAMS
Viking/Ecco:
2011. 448 pp.
£18.99/\$29.99

who associated the two in their cult of the war god Mars.

Other entertaining elementary experiments conducted by the author involve phosphorus and iodine. Doubting the claim that rotting herrings emit light, he left some decaying fish in his garage. Two nights later, he observed the phosphorescent "glowing of the lifeless herring", mentioned by twentieth-century writer

W. G. Sebald. Using local seaweed, Aldersey-Williams also prepared an intense violet vapour and black crystalline condensate of iodine. Here, he notes, he followed the poet Johann Wolfgang von Goethe, who in 1822 demonstrated iodine vapour and crystals for house guests in support of his controversial theory of colours.

For all its technical accuracy, *Periodic Tales* is neither a book of experiments nor a science book. Aldersey-Williams eschews the territory covered, for example, by Peter Atkins in *The Periodic Kingdom* (Basic Books, 1995). There are few references to atomic weight and atomic number, scarcely any chemical formulae and nothing on electrons and orbitals. There is no up-to-date periodic table among the quirky illustrations, merely the handwritten version created by Dmitri Mendeleev in 1869. Instead, the book is a cultural history of some of the chemical elements, dwelling on both their material presence in our lives and their figurative presence in art, literature, language, history and geography. Thus we come to know the elements individually, argues Aldersey-Williams, who regrets that his own formal chemistry education did so little to acknowledge such a "rich existence". So do I.

➔ **NATURE.COM**

For Nature's International Year of Chemistry special: go.nature.com/lyqnoij

J. L. AMOS/CORBIS

Aluminium, for example, was first named aluminum by Humphry Davy, who repeatedly tried to isolate it from its oxide, alumina. He followed the naming precedents set by platinum, molybdenum and tantalum. Then, in 1812, an anonymous reviewer of Davy's book *Elements of Chemical Philosophy* insisted on a name that sounded more "classical" — aluminium. Nonetheless, when use of the metal took off in the United States at the end of the nineteenth century, Americans plumped for the version that omits the letter 'i'. Not even the US literary critic H. L. Mencken could work out why in his 1919 book *The American Language*.

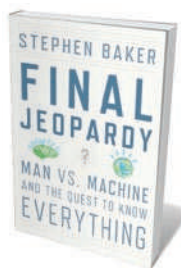
Aluminium enjoyed a brief mid-century vogue as a precious metal before the invention of the electrolytic separation process in 1886, still used today, which extracts it from bauxite (named after Les Baux in Provence, France, where the ore was found). In 1855, a French chemist, Henri Sainte-Claire Deville, managed to extract the metal by heating anhydrous aluminium chloride with sodium. But this was hugely expensive. His ingots — worth a dozen times more than silver — were admired at the Paris Universal Exposition of 1855 by Emperor Napoleon III of France, who offered financial support to Deville. Such was the metal's rarity that a renowned goldsmith, Christofle, made hand-crafted aluminium jewellery and tableware, which was favoured at imperial banquets, and an aluminium rattle was given to the emperor's newborn son. Chemical elements, *Periodic Tales* emphasizes, can go in and out of fashion. Think of what happened to chromium plating.

Almost every page yields a nugget. The difficulty, however, is to find order and meaning. Aldersey-Williams settles for five sections, divided into chapters on one element or a group such as the halogens. 'Power' includes gold, iron, carbon, plutonium and mercury; 'Fire' includes sulphur, phosphorus, chlorine, oxygen and radium; 'Craft' — tin, silver, copper, aluminium and calcium; 'Beauty' — chromium, arsenic, vanadium, antimony and neon; and finally 'Earth', encompassing the rare earth elements and some other, less familiar ones. This division is workable, but I query some choices. Gold, for instance, surely belongs as much in 'Craft' and in 'Beauty' as in 'Power'. It also seems odd to omit a chapter on silicon, given its starring role in electronics.

That said, the book is imaginative and fun. Who can resist the information that an unofficial Dutch spectroscopic analysis of the five-euro banknote shows it to be impregnated with an anti-counterfeiting ink containing a little-known rare earth element — europium. ■

Andrew Robinson is a writer based in London, and author of *The Story of Measurement*.
e-mail: andrew.robinson33@virgin.net

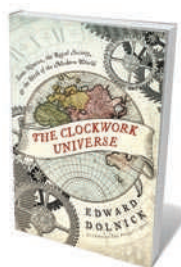
Books in brief



Final Jeopardy: Man vs. Machine and the Quest to Know Everything

Stephen Baker HOUGHTON MIFFLIN HARCOURT 288 pp. \$24 (2011)

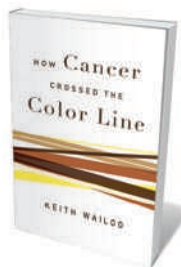
For the past year, IBM researchers have been building a robot clever enough to compete in the US television quiz show *Jeopardy!* In mid-February, viewers in the United States will be able to watch a real contest between man and machine, when two previous winners take on the drone. Technology writer Stephen Baker describes in his book how artificial-intelligence researchers constructed the robot and the challenges they faced in getting 'Watson' to understand language, spot puns and recall general knowledge.



The Clockwork Universe: Isaac Newton, the Royal Society, and the Birth of the Modern World

Edward Dolnick HARPER 400 pp. \$27.99 (2011)

From a modern perspective, seventeenth-century science can appear strange. Rational descriptions of a clockwork Universe sat happily beside belief in omens, alchemy and the devil. By portraying the lives and discoveries of Johannes Kepler, Galileo Galilei, Isaac Newton and Gottfried Leibniz, science writer Edward Dolnick fleshes out these contradictions in the thinking of the time. Emphasizing their social relationships and collaborations, he also brings to life the network of the Royal Society in London.



How Cancer Crossed the Color Line

Keith Wailoo OXFORD UNIVERSITY PRESS 264 pp. \$27.95 (2011)

Cancer awareness and treatment have a strong socio-political element. Attitudes to race have influenced cancer concerns throughout the twentieth century in the United States, finds historian Keith Wailoo in his study of medical, cultural and sociological factors around the illness. From being an affliction that was mainly associated with white women, cancer has crossed cultural boundaries. But race, class and gender issues linger, for example in reports of high rates of breast cancer in affluent parts of California and in the poor health outcomes for black men with prostate cancer.



Discoverers of the Universe: William and Caroline Herschel

Michael Hoskin PRINCETON UNIVERSITY PRESS 272 pp. \$29.95 (2011)

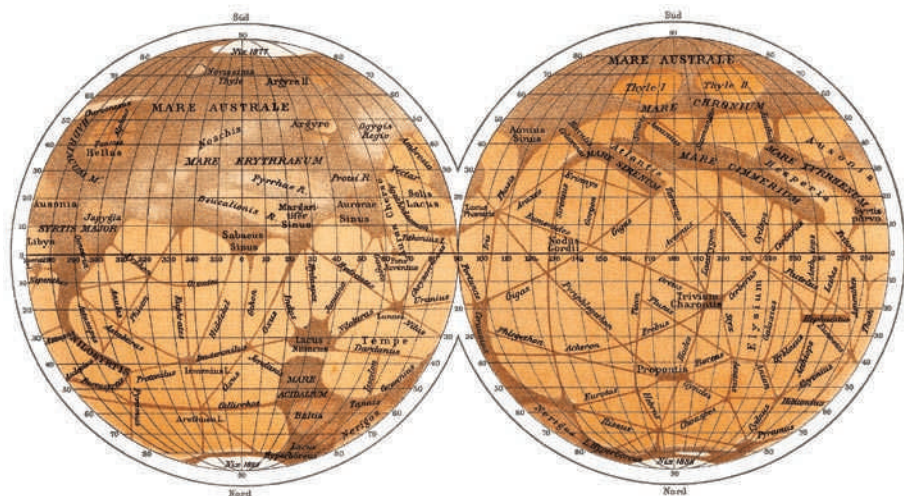
With the help of his sister Caroline, the eighteenth-century German-British astronomer William Herschel discovered the planet Uranus, revealed infrared radiation and coined the term asteroid. In this joint biography, written with the cooperation of the Herschel family, historian of astronomy Michael Hoskin portrays the siblings' shared passion for the night sky, and the triumphs and pitfalls of their work. Using an amateur telescope, the pair charted thousands of stars and nebulae in catalogues that are still used today. Caroline's role as one of the first professional women astronomers is also recognized.



Life in a Shell: A Physiologist's View of a Turtle

Donald C. Jackson HARVARD UNIVERSITY PRESS 192 pp. \$29.95 (2011)

Over 200 million years of existence, turtles have shared the planet with dinosaurs, witnessed the diversification of mammals and seen the spread of humans. Physiologist Donald Jackson conveys his love of the reptile in his book. He explains how its slow movements help it to survive winters under ice and describes how its shell functions as a home, armour and a buoyancy aid. By focusing on the physiology of this one familiar beast, he also reveals how scientific understanding evolves by building on the work of others.



Giovanni Schiaparelli's map of linear 'canals' on Mars sparked a debate that lasted for more than 80 years.

ASTRONOMY

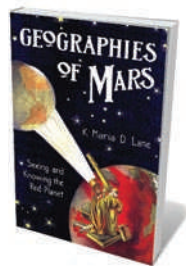
Martian illusions

The Mars canal controversy is a reminder to be cautious when interpreting alien worlds, notes **Michael Carr**.

Present-day Mars is dry, cold and inhospitable, yet we know from rovers and orbiting satellites that it has a rich geological and climate history. Past conditions may even have been benign enough to support forms of life. The possibility of life on Mars is, of course, not a new idea. In the late nineteenth and early twentieth century, the public was captivated by reports of canals on Mars, supposedly built by an advanced civilization in response to a desiccating planet.

Maria Lane's meticulously researched *Geographies of Mars* describes the canal controversy. She explains the intellectual and social factors that fed into the canal concept and its broad acceptance. The view of Mars as an "arid dying, irrigated world peopled by unfathomably advanced beings" grew from the geopolitics of European imperialism and US expansionism, Lane argues. Modern Mars is not discussed.

The basic story is familiar. Following the close opposition of Mars to Earth in 1877, Italian astronomer Giovanni Schiaparelli published a cylindrical-projection (Mercator) map of the red planet's surface. It showed numerous linear features, which he termed *canali*, that did not appear on other portrayals. The reality of the lineaments was initially questioned, mainly by British astronomers. But after their independent confirmation in 1886 by the European observers François Terby and Henri Perrotin, there was an explosion of canal sightings. By the end of the nineteenth century, most of the published



Geographies of Mars: Seeing and Knowing the Red Planet

K. MARIA D. LANE
University of Chicago
Press: 2010. 266 pp.
\$45

less, the canal idea died hard.

In 1961, French astronomer Audouin Dollfus published drawings showing canals, and in 1964, US astronomer Earl Slipher published photographs that he claimed removed any doubt about the canals' existence. Linear features were even portrayed on some of the charts that my colleagues and I used in 1971 during NASA's Mariner 9 Mars mission.

How did this come about? Lane suggests that presenting the canals in cartographic form gave them authority. The maps made implicit claims about the surface of Mars, conveying certainty that the same features would appear in the same place. Initially,

maps showed Mars's surface criss-crossed by a spider-web pattern of canals.

The US astronomer Percival Lowell added 116 new canals to Schiaparelli's map, and forcefully argued in highly publicized talks, books and magazine articles that the canals were built by intelligent beings. This 'sensation' sputtered out after 1910 as better photographs of Mars failed to reveal the features. Neverthe-

most observers simply published sketches of what they had seen. In 1878, English astronomer Nathaniel Green also published a Mercator map. Green's map was subtly shaded and lacked canals, whereas the features of Schiaparelli's were crisply defined. Owing to its clarity, Schiaparelli's map became accepted, as were his Mediterranean names for Martian landmarks.

Another factor in the acceptance of the canals was the superior observations claimed by those who promoted the canal idea. Lowell scorned observing conditions in Europe and the eastern United States, preferring to do his work in the western state of Arizona, with its mountains, dry air, isolation and environmental purity. He capitalized on contemporary enthusiasm for the wilderness, claiming that mountains were places of transcendence and divinity, sites of purity and vision. The view of astronomers as explorers conquering mountains and undergoing hardships for the cause of science was promoted, and comparisons were made with the polar expeditions of Robert Falcon Scott and Robert Peary.

Most sensational was Lowell's proposal that the canals were irrigation channels built by intelligent life, an idea that captivated public attention and provoked disagreement among scientists and commentators. By the 1890s, Mars was widely viewed as a vast desert, and its habitability was argued in that context. The most prominent debaters were Lowell and British biologist Alfred Russel Wallace.

Wallace claimed that the biological conditions necessary for life were not met on Mars. He noted that temperatures were unlikely to be warmer than on the Moon, and that there seemed to be little water. Lowell acknowledged that conditions were harsh but held the view that they were not severe enough to kill off all life. If one accepts Lowell's maps and their clearly artificial patterns as representing the truth, then his conclusions had some logic. But it is still a puzzle as to why Lowell and his followers became so convinced that they could see the spider-web patterns. Lane suggests that the inhabited Mars theory was also tied to the perceived objectivity of maps. When that objectivity faltered with the acquisition of better photography, so did belief in intelligent Martians.

Lane does not discuss the contemporary implications of this saga. An obvious analogy is the 'Face on Mars' controversy, in which a face-shaped hill seen in a poor-resolution image taken by the Viking 1 orbiter in 1976 was interpreted by some as evidence of an advanced civilization. Later, when images of much higher resolution showed that the hill was not face-shaped at all, a government conspiracy was invoked. Similarly, isolated hills have been interpreted as pyramids and surface streaks as runways.

Lane criticizes the process of naming features on the nineteenth-century Mars maps as

NATURE.COM

For a review of Paul Davies's book on the search for alien life: go.nature.com/milnuy

nationalistic manoeuvring. The astronomical community today is sensitive to this issue; for example, large river channels are now named using the word for Mars or star in various languages, and small craters are named after towns and villages from across the globe.

Professional astronomers have criticized Lowell for his cultivation of the media. Some planetary scientists today are similarly uneasy about the part that publicity plays in the Mars

exploration programme. Discoveries and their implications are kept confidential and announced with great fanfare at press conferences before being presented and challenged at scientific meetings. As a consequence, a more sensational interpretation of a newly discovered feature can get the most attention, irrespective of its merit.

After the Viking Mars landers failed to detect life in the late 1970s, geneticist Norman

Horowitz cautioned that the seductive idea that life could have started on Mars means we should take care in interpreting new findings and presenting them to the public. Lane's book reminds us that is still good advice. ■

Michael Carr is a planetary scientist with the US Geological Survey, Menlo Park, California, USA.
e-mail: carr@usgs.gov

SCULPTURE

The brain in a nutshell

Martin Kemp explains the resonances of Pascale Pollier's autopsy-inspired sculpture.

Medicine has long used visual representation in ambitious ways. This is particularly true if we include the illustrated herbals inspired by the great five-volume encyclopaedia by Dioscorides (AD 40–90). Because art has traditionally centred on issues of human existence, medicine has also inspired many artists.

Recent works based on medical themes have tended to use metaphor and allusion rather than direct illustration. A striking example is provided by Belgian biomedical artist and poet Pascale Pollier, in a sculpture currently on show in the exhibition *Picturing Science* at the Riverside Gallery in Richmond, near London.

Her intense piece is enigmatically entitled *Autopsy in a Nutshell*. A bell jar, into which two coils of wire enter, contains a magnifying glass, two light-emitting diodes and a jointed stand with two sprung clamps. The beaks of the clamps grip a small model of the human brain and half of a walnut shell, the inside of which has been minutely remodelled to match the inside of a cranium.

As the title suggests, it was inspired by Pollier's witnessing of an autopsy. The first version of the piece was commissioned by Belgian learning expert Bernard Lernout, a great aficionado of Leonardo da Vinci and a fan of Michael Gelb's historically eccentric but creatively ingenious book, *How to Think Like Leonardo da Vinci* (Delacorte Press, 1998). Lernout directed Pollier to Gelb's seven Leonardesque "principles": curiosity (*curiosità*), demonstration (*dimostrazione*), sensation (*sensazione*), smokiness or ambiguity (*sfumato*, a layered paint effect), art-science (*arte/scienza*), embodiment (*corporalità*) and the connections between things (*connessione*).

Pollier picked up on three of these: demonstration, defined by Gelb as "learning



Autopsy in a Nutshell (2006) exploits more than the visual similarity between the brain and a walnut, as revealed on closer inspection (bottom).

Picturing Science

Riverside Gallery,
Richmond, UK.
Until 26 February
2011.

from experience"; art-science, as balancing the properties of the two sides of the brain; and connection, as

the need to see the big, linked-up picture. Her modestly sized, elaborate and detailed construction does not illustrate an autopsy, rather its making is framed by Gelb's three principles. Her artwork invites us to read meaning into the conjunction of objects. Faced with an image as powerful as that of a brain removed from its bony container, we can take up her invitation.

But why the walnut? It clearly exploits the visual resonance between a furrowed walnut plucked whole from its halved shell and the wrinkled configuration of the brain. It also refers to the ancient and cross-cultural idea of the microcosm and macrocosm, which highlights similarities of form and function across every scale in nature and the wider Universe. Old herbal medicine in both Western and Eastern cultures used this doctrine to help determine the source of treatments. A herb or fruit that resembles a human organ was seen as potentially efficacious for treating a disease of that organ.

Before we smile patronizingly at such ancient mysticism, it is curious to note that walnuts could have an effect on some ageing disorders of the brain. The late James Joseph and his team at Tufts University in Boston, Massachusetts, reported in the *British Journal of Nutrition* in 2009 that a diet including walnuts seemed to improve cognitive function in ageing rats.

As happens in the best scientifically orientated artworks, a visual starting point opens up a range of associations across historical and contemporary practice. ■

Martin Kemp is emeritus professor of the history of art at the University of Oxford, UK.

CORRESPONDENCE

Pick sanitation over vaccination in Haiti

I have been investigating Haiti's water system since 2007 and strongly believe that the limited resources available to combat the country's cholera epidemic should be spent on sanitation and clean water, rather than on vaccination (*Nature* **469**, 273–274; 2011). Otherwise, the local geology and ecology will allow cholera and other water-borne pathogens to persist.

Haiti has a backbone of granitic igneous rocks near the border with the Dominican Republic, surrounded by sedimentary limestone and shale. Fissures in limestone give rise to shallow aquifers that are especially prone to contamination by water-borne pathogens.

The devastation of last year's earthquake in Haiti joined intractable problems of poverty, deforestation and the loss of its microbiotic ecosystem. Soil microorganisms that consume pathogens are integral to the macrobiotic ecosystem, and are a first line of defence against groundwater contamination. The loss of Haiti's soils and the beneficial organisms they host means that many shallow aquifers are now unprotected.

Pathogens that thrive in Haiti's warm groundwater are flushed out by heavy rains and hurricanes, helping them to spread and cause new disease outbreaks. This is why funds should be spent on long-term, sustainable and resilient water resources.

Peter Wampler *Grand Valley State University, USA.*
wamplerp@gvsu.edu

Harnessing value of dispersed critiques

You mention the 'publish and be damned' model in academia, in which scientific work is

disseminated, regardless of merit, for rapid and public online criticism rather than slow, private peer review (*Nature* **469**, 286–287; 2011). For this to work, we must devise new ways to link widespread Internet discussions to and from an original paper.

Under current academic publishing models, one could easily miss substantive reactions to a paper that appear in other peer-reviewed journals. Even when a paper is retracted, research shows that this information is poorly disseminated and that the paper can continue to be cited widely and positively for years afterwards (K. M. Korpela *Curr. Med. Res. Opin.* **26**, 843–847; 2010).

The growth of blogs, Twitter and free online access have caused a welcome explosion in scientific content. But this is atomized and interconnected by a hotchpotch of linking and referencing conventions. If we are going to harness its true value, we shall need dedicated librarians and information scientists to find ways of automating the process of linking content together again. That in itself would be a transgressive scientific innovation.

Ben Goldacre *London School of Hygiene and Tropical Medicine, UK.*
ben.goldacre@lshtm.ac.uk

Biomarkers: better donor protection

George Poste calls for the creation of international biobanks as part of research efforts on disease and drug-response biomarkers (*Nature* **469**, 156–157; 2011). As the director of Israel's biobank and a member of international biobank organizations, I must point out that this can only work if the use of genetic information is guaranteed to be non-discriminatory.

Laws that guard against such discrimination are essential for public trust in biomedical research and to protect the identity of donors, but few countries have laws in place that are sufficiently comprehensive — including the United States (R. Korobkin and R. Rajkumar *N. Engl. J. Med.* **359**, 335–337; 2008).

The need to safeguard genetic information is becoming more urgent. For example, personal sequence data sent over the Internet by direct-to-consumer providers are insufficiently protected; and biobank donors or customers who are protected in their home country may still face discrimination by employers or insurers in another. Such concerns risk discouraging potential donors and will hinder international biobanking efforts.

Calls to issue a genetic information non-discrimination amendment to the Helsinki Declaration of Human Rights have been voiced for some time (J. Harris and J. Sulston *Nature Rev. Genet.* **5**, 796–800; 2004). That might not solve all biobanking problems, but it could improve the international situation. The grim statistics on the poor yield of clinically valuable biomarkers serve as a sober reminder that the time has come for such an amendment.

David Gurwitz *Tel-Aviv University, Israel.*
gurwitz@post.tau.ac.il

Biomarkers: call on industry to share

It will be challenging to mobilize public funding for huge standardized repositories of biological specimens and accompanying clinical data (*Nature* **469**, 156–157; 2011). But this could be complemented by tapping into ongoing industry-sponsored

biobanking activities.

For example, clinical trials often include dedicated biomarker studies, with a budget for collecting specimens and data — a facility that could contribute to shared repositories, serving both industry and the public at a modest additional public cost. Even if only a fraction of the more than four million subjects enrolled worldwide in interventional trials were to be captured, progress would be enormous.

From my work on biomarker detection at biotechnology company SDIX, and as a consultant on biobanking for public-health authorities, it is clear that this expansion could also advance public health, providing biomarkers to improve our understanding of risk factors associated with population-specific diseases and helping us to tailor public-health initiatives.

Such a synergistic venture could significantly advance the field, but would require systematic encouragement from national health ministries, harmonization of protocols, integration of individual repositories into larger virtual networks, attention to ethical, legal and social implications, and respect for the need of private partners to retain certain intellectual-property rights.

Klaus Lindpaintner *Strategic Diagnostics (SDIX), USA.*
klindpaintner@sdix.com

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines at **http://go.nature.com/cmchno**. Readers are also welcome to comment online on anything published in *Nature*: **www.nature.com/nature**.

Jack Oliver

(1923–2011)

Seismologist who helped demonstrate that Earth's continents move constantly.

John “Jack” Ertle Oliver, who died on 5 January aged 87, was one of the founding fathers of modern seismology, plate tectonics and deep imaging of Earth's continental crust. He and Bryan Isacks, a former graduate student whom he had advised, showed that not just Earth's crust, but the entire 100-kilometre-thick outer layer of Earth, the lithosphere, moves over the face of the planet before plunging into the weaker underlying asthenosphere. This discovery gave the then-controversial theory of continental drift a solid foundation, paving the way for plate tectonics.

Two strong personalities had key roles in Oliver's scientific development. The first was Paul Brown, his high-school American-football coach in Massillon, Ohio — who later found fame as coach of the Cleveland Browns and founder of the Cincinnati Bengals. Brown was famously intolerant of prima donnas, and Oliver's philosophy of science and administration followed the same principle — that good scientists who worked together could produce more important science than the best scientists working alone.

Supported by a football scholarship, Oliver studied physics at Columbia University in New York, interrupted by three years of service in the Navy from 1943 to 1946. On returning he met Maurice Ewing, who became his PhD adviser. Ewing instilled in Oliver the view that going where no one had gone before offered a high probability of scientific discovery, and that discovery was fun — jet fuel to a man with insatiable curiosity.

In the 1950s, most seismologists recorded Earth's tremors by measuring the arrival times of two types of seismic wave, P and S waves, at their seismographs. These waves travel through the interior of Earth. Long-period waves that ripple over the surface of the planet offered a ‘place’ where few had gone before. Oliver was the first to recognize many of the peculiarities of the waveforms that computers now analyse and simulate. His papers from this period show exquisite recordings of surface waves, buttressed by sensible interpretation.

Towards the end of that decade, Oliver joined about a dozen of the country's outstanding scientists to form the Berkner Panel on Seismic Improvement. This was meant to

raise the game of seismology in the hope of finding ways to detect and identify underground nuclear explosions. (The Limited Test Ban Treaty, which banned nuclear tests from the atmosphere, underwater and in space, but not from underground, was ratified in 1963.) Modern seismology grew rapidly out of the panel's recommendation to build a global suite of detector stations — the World-Wide Standardized Seismograph Network.

Following Ewing's philosophy again, Oliver and Isacks installed seismographs in Fiji and Tonga to study what were then

not occur where temperatures were too great to support the needed stresses, but instead happened within slabs of still-cold lithosphere that had plunged rapidly to great depth.

Like many founders of plate-tectonics theory, Oliver soon changed direction. In 1971, he left Columbia University to revive the geological-sciences department at Cornell University in Ithaca, New York, and turned his attention to one of the blind spots in Earth science, the lower crust. He launched the Consortium for Continental Reflection Profiling, which pioneered the use of seismic

reflection technology — developed in the oil industry — to explore the lower crust of continents. It has since been copied by numerous nations.

He always worked to instil a sense of teamwork, and tried to erase the differences between geologists (sometimes belittled for being quantitatively challenged) and geophysicists (sometimes perceived as believing only what they could not see with their eyes). Cornell led the widespread transformation of geology and geophysics departments into integrated Earth-science departments.

Oliver continually inspired students around him, asking, almost as an interrogative mantra: “What is the next most important problem?” He once said, “Sure. I want to work with bright students, but what I really want are students who can ask good questions.” Time and again he demonstrated that asking the right question was a shortcut to answering an important one. Spicing his later writings with limericks, he once wrote:

*“If creativity is what you strive for,
The status quo you must learn to abhor,
Chains of convention unfetter,
Seek the different yet better,
Pay no attention to those keeping score!”*

Not a self-promoter, Oliver was as at home with technicians as with academicians. He died while in lucid conversation with his long-time administrative assistant, Judy Healey, abruptly announcing, without concern: “I'm done.” ■

Peter Molnar is at the Department of Geological Sciences at the University of Colorado in Boulder, Colorado 80309, USA. e-mail: peter.molnar@colorado.edu



one of Earth's enigmas: deep earthquakes. These occur at 300–700 kilometres below the surface, where pressures and temperatures seemed far too high to allow rock to fracture. With seemingly little more than a glance at their seismograms, they knew they had something important: P and S waves with unusually high frequencies. They realized that the lithosphere (which includes the planet's crust and the uppermost ‘cool’ part of the mantle) underlying the Pacific Ocean to the east had plunged to a depth of 700 kilometres. Much of the scientific community was slow to appreciate their insight, for few seismologists had turned their attention to the theory of continental drift, or even taken it seriously, but scepticism quickly gave way to consensus with another of their studies.

In 1968, with Lynn Sykes, another researcher who had been advised by Oliver, Oliver and Isacks published what must be the most widely read seismological paper ever written, ‘Seismology and the new global tectonics’ in the *Journal of Geophysical Research*. Among their many insights, they realized that the formerly enigmatic deep earthquakes did

Squinting at quantum systems

Quantum measurements always have a back-action: they ‘kick’ the system in a particular way. This can be used to drive the system to any desired state using a fixed type of measurement, provided it can be ‘unsharpened’.

HOWARD M. WISEMAN

For the purpose of controlling a system, two facts appear self-evident. First, the more information one can obtain about the system, the better one can control it. Second, one needs to do more than just obtain information in order to control the system. In the quantum world, however, self-evidence cannot be trusted. Writing in *Physical Review A*, Ashhab and Nori¹ refute the two ‘facts’ just given and show that a quantum system can be quickly driven to any desired state using a fixed type of measurement. Although various schemes have been proposed^{2,3} for driving a quantum system from one state to another using only quantum measurements, this is the first time it has been shown to be achievable using repetitions of a given measurement. Crucially, the authors’ proposal requires the measurement to be unsharp. That is, one must avoid obtaining too much information about the system.

Driving a system to a desired target state is a common control problem in physics and engineering. An everyday example is driving your car to a desired location. In practice, this involves observing what is happening all the time, but the observation is not what drives the car. In principle, if you knew your initial state (location), and there was no ‘noise’ (no other cars on the road), and you had a good theory (a perfectly memorized route map, and a knowledge of how your car responds to its controls), then you could drive your system (car) to the desired state with no observation whatsoever (with your eyes closed).

In quantum physics, observation can have a much more active role than it does in driving a car. This is related to Heisenberg’s uncertainty principle, which implies that even if the system is in a pure state (that is, we know as much about the system as nature allows), some of its properties must be unpredictable. As a consequence, making a certain type of measurement of a system to reveal more about a certain property (for example, its momentum) will necessarily disturb other properties (for example, its position), making them more uncertain. This happens even if the measurement is minimally disturbing⁴, so that it leaves the system in a pure state as close as

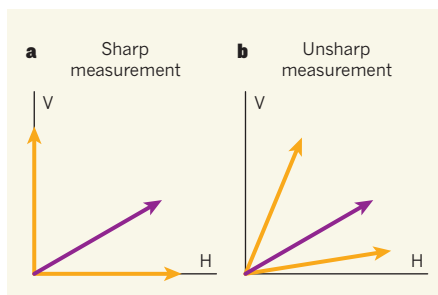


Figure 1 | Sharp and unsharp measurements.

The purple arrow represents the initial pure quantum state of a two-dimensional quantum system such as the polarization of a photon. For this physical system, the direction H represents a horizontally polarized state and V represents a vertically polarized state, and the initial state is a particular superposition of these (but closer to H than V). **a**, After a sharp measurement of polarization, the photon is in either the H-polarized state or the V-polarized state (orange arrows). **b**, After an unsharp measurement (‘squinting’; here with a sharpness of 50%), the photon is still in a superposition, but the ‘H’ result puts the state (lower orange arrow) much closer to an H-polarized state than the ‘V’ result puts it to a V-polarized state (upper orange arrow). Ashhab and Nori¹ show that by tweaking the sharpness of the measurement to the right value, the measurement-induced change in the state can be optimized.

possible to the initial pure state, on average⁵.

The unavoidable disturbance necessitated by gaining information about a quantum system is known as quantum back-action, and the change it causes in the system’s state can be used to control the system. In particular, probing a system — even in a minimally disturbing way — can replace the application of direct controls to drive it into a desired target state^{1–3}. What sets the proposal of Ashhab and Nori¹ apart is its innovative use of measurement strength, or ‘sharpness’⁴.

In traditional quantum mechanics, one considers sharp measurements⁴, in which the final pure state is determined solely by the measurement outcome (that is, it is independent of the initial state). If one could design the measurement so that the set of possible final states included the desired target, then clearly one would have the possibility of causing the

system to jump straight there by a single measurement. But what if these possible final states are fixed, and do not include the desired state?

What Ashhab and Nori have shown is that for one class of measurements (‘symmetric, informationally complete’ measurements⁶) any target state can be attained, as long as the measurement can be repeated, and as long as one can make it unsharp⁴. As would be expected, unsharp measurements have less quantum back-action than sharp measurements (Fig. 1). They cause jumps towards — rather than directly to — the possible final states of the corresponding sharp measurement. Unsharp measurements are known to be better than sharp measurements in feedback control of quantum systems^{7,8}. Ashhab and Nori¹ introduce a completely new sort of feedback, in which the degree of sharpness for each measurement depends on the outcomes of previous measurements.

To return to the analogy of driving a car, in quantum physics it is as if your car would jump to a different location every time you opened your eyes to observe it. For a sharp observation (eyes wide open), there would be only a fixed set of possible post-jump locations, so if none of these was your target location, then you would never get to where you want to go. But with an unsharp measurement (squinting), how far the car would jump towards one of those fixed locations would depend on the car’s current location. By squinting just the right amount each time, you would have a good chance of reaching your target location after only three observations. This is what Ashhab and Nori have shown to be possible using unsharp measurements for a two-dimensional quantum system (such as the polarization of a photon).

This work raises a number of obvious questions. Can one prove analytically what Ashhab and Nori have shown by numerical calculations? How does it generalize to higher-dimensional systems? Is having an informationally complete⁶ measurement sufficient? Squinting at a quantum system may not be the easiest way to drive it to a target state, but answering these questions will continue to reveal more about the fascinating and non-self-evident field of quantum measurement and control. ■

Howard M. Wiseman is at the Centre for Quantum Computation and Communication Technology (Australian Research Council), and the Centre for Quantum Dynamics, Griffith University, Nathan, Queensland 4111, Australia.
e-mail: h.wiseman@griffith.edu.au

1. Ashhab, S. & Nori, F. *Phys. Rev. A* **82**, 062103 (2010).
2. Pechen, A., Il'in, N., Shuang, F. & Rabitz, H.

- Phys. Rev. A* **74**, 052102 (2006).
3. Jacobs, K. *New J. Phys.* **12**, 043005 (2010).
4. Wiseman, H. M. & Milburn, G. J. *Quantum Measurement and Control* (Cambridge Univ. Press, 2010).
5. Banaszek, K. *Phys. Rev. Lett.* **86**, 1366–1369 (2001).
6. Renes, J. M., Blume-Kohout, R., Scott, A. J. & Caves, C. M. *J. Math. Phys.* **45**, 2171–2180 (2004).
7. Wiseman, H. M. & Thomsen, L. K. *Phys. Rev. Lett.* **86**, 1143–1147 (2001).
8. Gillett, G. *et al. Phys. Rev. Lett.* **104**, 080503 (2010).

ANIMAL BEHAVIOUR

The nexus of sex and violence

In mice, brain neurons that respond during either mating or aggression exhibit spatial overlap, and some even respond during both. This may help to explain the relationship between sex and violence in human behaviour. [SEE ARTICLE P.221](#)

CLIFFORD B. SAPER

The close relationship between sex and violence has been an enduring theme in literature, theatre and music since the dawn of 'civilized' culture. A particularly graphic depiction of the connection can be found in Anthony Burgess's book *A Clockwork Orange*, which famously mixes the two in a *mélange* of 'ultraviolence'. Although fascinating, the intertwined nature of these two opposites of social interaction and the underlying neurobiological basis have remained a puzzle. On page 221 of this issue, Lin and colleagues¹ identify some of the basic circuitry for these behaviours in the hypothalamus — a primitive part of the brain that has been highly conserved throughout mammalian evolution.

It has long been known² that, in cats, electrical stimulation in certain regions of the hypothalamus elicits attack behaviour. Studies^{3,4} in rats have also identified a network of brain sites in which stimulation can produce aggression, including the ventromedial nucleus of the hypothalamus (VMH).

Meanwhile, investigators have identified⁵ VMH neurons that express receptors for sex hormones, and have shown⁶ that electrical stimulation of the VMH can produce sexual behaviours in rats. Moreover, after mating or aggression, neurons in the ventrolateral VMH (VMHvl) and several other brain areas, including parts of the amygdala, express cFos — a protein that is expressed by many brain neurons that have recently undergone activation⁷. It remained unclear, however, whether both behaviours activate the same neurons, or separate cell populations that overlap spatially. Lin and co-workers¹ attempted to sort this out.

The authors sequentially exposed male mice to another male and, 15–20 minutes later, to a female — situations that would trigger first aggression and then sexual behaviour. They then analysed neurons for two types of cFos messenger RNA: heteronuclear mRNA, which would have been produced more recently, while the female mouse was present; and cytoplasmic mRNA, which would have matured from heteronuclear mRNA generated earlier while the male mouse was present. Although neurons expressing both types of mRNA spatially overlapped in the VMHvl, they largely belonged to distinct populations. However, a proportion (20–30%) of these cells showed cFos expression during both encounters.

To better define the time course of the neuronal response, Lin *et al.* recorded the firing of individual VMHvl neurons in male mice during encounters with both sexes. Whereas some 40% of the VMHvl neurons were excited by male intruders, about half of these were activated only during close encounter and attack. By contrast, roughly one-third of the VMHvl cells were excited by a female intruder, but the level of excitement in around two-thirds of these neurons tended to decrease as the sexual encounter progressed.

About half of all recorded neurons in the VMHvl responded initially to both a male and a female intruder, but many of these ultimately continued firing during only one of the two behaviour patterns. The dual activation of some neurons during the earliest stages of both encounters indicates that they share some types of input; in other words, the interaction of the two outcomes is deeply rooted in the basic architecture of the brain.

Lin *et al.*¹ provide another line of evidence for the interaction between sexual behaviour



50 Years Ago

The Tobacco Manufacturers' Standing Committee has as a declared aim the assistance of research into questions concerned with the relationship between smoking and health. That this object is being fulfilled is evident from its report for the year ended May 31, 1960, which summarizes investigations carried out during the year under the auspices of the Committee or with its financial support ... Fractions of cigarette smoke condensate prepared in the laboratories of the Committee have been found by several workers to have carcinogenic or tumour-promoting properties, but as the report points out, these results, obtained by application of smoke fractions to animal tissues, are not necessarily reliable guides to the possible response of human lung tissue to tobacco smoke.

From *Nature* 11 February 1961

100 Years Ago

The terrible intensity of the outbreak of pneumonic plague now raging in Manchuria, and the presence of plague-infested animals within our own borders, have called forth recently a number of communications on plague in the daily press. A special correspondent in *The Times*, in two well-informed articles ... summarises the situation, and gives an admirable sketch of the principal facts concerning the modes of spread of plague. Dr. L. W. Sambon has also contributed two letters on the subject ... He remarks, for example, that in his belief transmission from man to man is probably more frequent than from rat to man. If Dr. Sambon bases this statement upon personal experience of epidemics of bubonic plague, it must be said that his observations are directly opposed to the experience of many competent plague workers.

From *Nature* 9 February 1911

and violence by activating the VMHvl — using the technique of optogenetic stimulation — during encounters with intruders. The male mice rapidly attacked animals of either sex; most would even attack an inflated glove if it was moved. Intriguingly, however, when the authors stimulated the VMHvl at the same level during sexual intercourse, they saw no attack. This suggests that neurons mediating aggression are actively suppressed during mating, even in mice that are made hyperaggressive by optogenetic stimulation.

Can hypothalamic neurons be manipulated to curb aggressive behaviour? Lin and colleagues used a viral vector to cause VMHvl neurons to express inhibitory chloride channels that are regulated by ivermectin — an antibiotic that can be given systemically and that penetrates the brain. Following ivermectin administration, 25% of animals that previously showed normal attack rates did not attack at all, and the rest showed increased latency and decreased duration of attacks on male intruders. Eight days after ivermectin injection, the responses had returned to normal levels.

Clearly, much more must be learned about

how hypothalamic neurons that mediate sex and violence operate, and how they can be productively controlled. But the possibility of using this information to change human behaviour may not be far from the minds of those who live with the challenge of dealing with violent sexual offenders. In *A Clockwork Orange*, social engineers used Pavlovian conditioning of the protagonist to induce an aversion to both sex and violence. Could — and should — the behaviour of sex offenders or violent criminals be similarly controlled by genetic therapy to induce expression of ion channels in the VMHvl followed by pharmacological or optogenetic stimulation? It would be particularly valuable to decipher the chemical features of neurons with specific firing patterns. With this information in hand, researchers could potentially design vectors to introduce foreign ion channels only in a specific group of neurons — a way to differentially modify sexual or violent behaviours.

It is noteworthy that this work¹ was done entirely in male mice. Although among humans, men commit a larger proportion of both sex offences and violent crimes,

women commit their own distinct patterns of sexual and violent crimes. The same parts of the brain exist in both sexes, and presumably similar circuitry controls behaviour in females. Exploring sexual differentiation of the circuitry that regulates sex and violence might provide an exciting chapter in understanding some of the most basic components of our personalities. ■

Clifford B. Saper is in the Department of Neurology, Program in Neuroscience, and Division of Sleep Medicine, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA.
e-mail: csaper@bidmc.harvard.edu

1. Lin, D. *et al.* *Nature* **470**, 221–226 (2011).
2. Hess, W. R. & Akert, K. *Arch. Neurol. Psychiatry* **73**, 127–129 (1955).
3. Chi, C. C. & Flynn, J. P. *Brain Res.* **35**, 49–66 (1971).
4. Lammers, J. H. C. M., Kruk, M. R., Meelis, W. & van der Poel, A. M. *Brain Res.* **449**, 311–327 (1988).
5. Schleicher, G., Stumpf, W. E., Morin, J. K. & Drews, U. *Brain Res.* **397**, 290–296 (1986).
6. Pfaff, D. W. & Sakuma, Y. *J. Physiol. (Lond.)* **288**, 189–202 (1979).
7. Veening, J. G. *et al.* *Eur. J. Pharmacol.* **526**, 226–239 (2005).

PALAEOCLIMATOLOGY

Core data from the Antarctic margin

Sediments at the edge of Antarctica are a largely unexploited source of information about climate change. They have now provided a valuable local record of sea surface temperatures for the past 12,000 years. [SEE LETTER P.250](#)

JAMES BENDLE

The coastal areas of Antarctica are sites of strong air–sea–ice interactions that can affect the entire globe, but they remain the least-studied region on Earth with respect to climate variability. The paper by Shevenell *et al.* (page 250 of this issue¹) is a good start in addressing that deficiency.

The western Antarctic Peninsula is warming five times faster (3.4°C per century²) than the global mean increase during the twentieth century. The consequences are evident in the changing distribution of plants and animals, and in the retreat and sometimes dramatic disintegration of ice shelves. But there are all too few data to assess the causes of this warming, and to judge how unusual it is in recent Earth history. Satellites have been continuously monitoring sea surface temperatures (SSTs) and ice extent on the Antarctic Peninsula for only the past three decades. Globally, the instrumental record covers just the past few hundred years, with notably sparse data coverage

for the Southern Ocean and Antarctica.

It is against this background that the paper by Shevenell *et al.*¹ appears. The authors report a proxy record of SSTs on the inner continental shelf of the western Antarctic Peninsula for the Holocene epoch — the most recent, relatively warm and stable 12,000 years of geological time. They reconstruct SSTs using measurements of biological molecules extracted from ocean-sediment cores collected by the Ocean Drilling Program (ODP Leg 178), from Site 1098, at 1,010 metres water depth in the Palmer Deep, a prominent basin on the inner shelf. Figure 1a of the paper (page 251) shows the region concerned.

Over the interval 12,000–2,000 years ago, the reconstructed SSTs exhibit a cooling trend of 3–4°C, which broadly tracks the decline of spring solar radiation at 65°S (a function of cyclical changes in Earth's orbit). Superimposed on the longer trend are temperature variations of 2–4°C on the centennial to millennial scales. The longer-term cooling trend and many of the millennial-scale events can be found in

other palaeoclimate records³, but in comparison with other locations the temperature variations are remarkably high.

Thus, the absolute values should be treated with caution, but perhaps are not unexpected. Because of its position in the seasonal sea-ice zone, models predict that temperature changes at ODP 1098 should be amplified by the albedo effect (for example, dark, sunlight-absorbing open water warms up more readily than white, sunlight-reflecting sea ice)⁴. The temperatures between 11,800 and 9,000 years ago are especially warm and may also reflect intense spring/summer warming of surface waters, stratified by glacial meltwater. Interestingly, independent geological evidence indicates that the neighbouring George VI Ice Shelf collapsed around 9,600 years ago, following the 2,000 years of warm temperatures recorded at ODP 1098.

Most intriguingly, Shevenell and colleagues' work supports evidence that, during the Holocene, SSTs off the western Antarctic Peninsula were directly linked to westerly wind strengths in the Southern Hemisphere and to the El Niño–Southern Oscillation (ENSO; a roughly periodic, trans-Pacific pattern of climate fluctuation). A clearer mechanistic understanding of these connections will be essential. ENSO and the southern westerlies are predicted to strengthen further with future climate warming; if they are indeed a strong controlling factor on the temperature of the oceans around Antarctica, there are implications for ice-sheet stability and sea-level changes.

But what about the recent rapid warming of the western Antarctic Peninsula? What has the

new record to tell us about that? Unfortunately, the piston-coring technology used to recover long sediment cores often scrambles the sediment–water interface, and the most recent decades in the ODP 1098 sequence may be missing. So, to complete the Holocene record, Shevenell *et al.* used temperatures estimated from samples collected nearby using different coring techniques. Those samples indicate a recent warming of about 3.2 °C, comparable to the observed trend of 3.4 °C per century.

Two factors were essential to Shevenell and colleagues' success. One was the core site itself; the other was the proxy 'biomarker' measurements they used to extract the temperature estimates.

Apart from the logistical challenge of working in Antarctic waters, the dynamic inner-shelf environment tends not to be conducive to the stable accumulation of sediments. But the Palmer Deep has been relatively undisturbed, and for 12,000 years the remains of planktonic organisms have settled there, leaving a temperature record for scientists to interpret. The 43-metre thickness of the ODP 1098 Holocene sequence is impressive, resulting from rapid deposition due to the local high productivity of plankton in the surface waters and the focusing of sediments into the basin. The record is well dated too, by 51 radiocarbon analyses on organic matter and calcite microfossils⁵.

As to the proxy Shevenell *et al.* used for temperature estimates, the enduring problem has been that established methods of reconstructing SSTs are often not feasible with Antarctic sediment cores. Chemical conditions at ODP 1098 mean that the calcium carbonate remains of planktonic foraminifera, traditionally employed for studying past SSTs, are almost entirely absent from the core record. An approach that exploits the alkenone biomarker, which is preserved in sediments in many parts of the world, was likewise not possible. The main marine alkenone producer, the coccolithophore *Emiliania huxleyi*, is relatively rare in Antarctic waters, and alkenones were not detected in the ODP 1098 record.

So the authors turned to the increasingly deployed TEX₈₆ index (tetraether index of tetraethers with 86 carbon atoms). The index uses the relative abundance of another biomarker, membrane lipids produced by marine archaea found in the open sea⁶. To maintain membrane viability, these archaea adjust their lipid composition according to the water temperature, and that record is preserved in sediments.

Mindful of the reported scatter in the relationship of TEX₈₆ to temperature in the surface sediments of polar regions⁷, Shevenell *et al.* analysed additional regional surface-sediment samples and derived an amended calibration for application at ODP 1098. They argue that their down-core temperature reconstruction is weighted towards spring (rather than mean

annual) temperatures. This argument is based on two considerations: first, on local ecological studies suggesting that living archaea are most abundant in the depth range 0–150 metres in the austral spring⁸; second, on the down-core similarities between the greatest abundances of the archaeal biomarkers and the resting spores of a diatom alga (*Chaetoceros*) that tends to bloom in the early spring.

Much remains to be done to improve our understanding of archaeal ecology, and of the seasonal and depth-integrated generation of the TEX₈₆ signal in Antarctic waters and how it is transferred to sediments. However, Shevenell *et al.*¹ demonstrate the exciting potential of applying biomarker proxies to high-resolution sediment archives from the Antarctic margins. More is to come. Workers on Integrated ODP Leg 318 recently recovered an annually laminated, 180-metre-thick Holocene sequence from the East Antarctic Adélie Basin⁹, which

should provide records of ultra-high — annual to decadal — resolution. ■

James Bendle is in the School of Geographical and Earth Sciences, University of Glasgow, Glasgow G12 8QQ, UK.

e-mail: james.bendle@ges.gla.ac.uk

1. Shevenell, A. E., Ingalls, A. E., Domack, E. W. & Kelly, C. *Nature* **470**, 250–254 (2011).
2. Vaughan, D. G. *et al.* *Clim. Change* **60**, 243–274 (2003).
3. Bentley, M. J. *et al.* *Holocene* **19**, 51–69 (2009).
4. Renssen, H., Goosse, H., Fichefet, T., Masson-Delmotte, V. & Kog, N. *Holocene* **15**, 951–964 (2005).
5. Domack, E. *et al.* *Holocene* **11**, 1–9 (2001).
6. Schouten, S., Hopmans, E. C., Schefuss, E. & Damsté, J. S. *Earth Planet. Sci. Lett.* **204**, 265–274 (2002).
7. Kim, J.-H. *et al.* *Geochim. Cosmochim. Acta* **74**, 4639–4654 (2010).
8. Murray, A. E. *et al.* *Appl. Environ. Microbiol.* **64**, 2585–2595 (1998).
9. Expedition 318 Scientists. IODP Prelim. Rep. 318, doi:10.2204/iodp.pr.318.2010 (2010).

STRUCTURAL BIOLOGY

A new look for the APC

Solving the structure of protein complexes is particularly challenging when they contain many subunits. In the case of the APC, a fruitful strategy has been to gain information by subtracting subunits. SEE ARTICLE P.227 AND LETTER P.274

IAN FOE & DAVID TOCZYSKI

Many cell-cycle regulators are targeted for degradation by being tagged with chains of the small protein ubiquitin. Ubiquitin ligase enzymes mediate the selection of target proteins. One of the most complex ubiquitin ligases is the anaphase-promoting complex (APC) — a 1.5-megadalton assemblage of 1–2 copies of each of 13 different subunits, as well as two diffusible activators. Despite extensive genetic and biochemical work on the APC, in-depth mechanistic studies of this complex have proved technically difficult. Three papers (two in this issue^{1,2}, and one in *Nature Structural & Molecular Biology*³) provide insight into how the APC mediates ubiquitination of its substrates. The researchers report high-resolution structures of the APC that they obtained by electron microscopy (EM), generated from structures with different combinations of subunits, allowing them to infer the exact placement of many of its subunits.

Previous studies^{4,5} had already created a blueprint of the APC's architecture (Fig. 1). Each APC consists of a platform (made up of the proteins Apc1, Apc4 and Apc5) to which two subcomplexes attach. The first of these, the TPR subcomplex, consists of homodimers of three — or, in metazoans, four — subunits (Cdc23, Cdc16 and Cdc27), which contain

tetratricopeptide repeats (TPRs). On electron micrographs, this subcomplex appears as an 'arc lamp', in which the Cdc23 homodimer serves as the contact point with the platform region. The TPR subcomplex also contains several small, non-essential subunits. The second subcomplex, known as the catalytic subcomplex, contains a cullin (Apc2) and a RING (Apc11) protein, which are common to the class of ubiquitin ligases to which the APC belongs. It also contains Doc1 — a subunit implicated in substrate binding.

Schreiber *et al.*¹ (page 227) devised an expression system producing active yeast APC by co-expressing subunits from its various stable substructures in single baculoviruses. They were thus able to generate APC subcomplexes and solve their structures individually using EM. The authors then compared structures containing various subunits with those lacking them, attributing the density missing from one structure to the missing subunits. They further used the high-resolution cryo-EM structure (10 Å) of the complex^{1,2} to 'dock' the crystal structures of individual APC subunits. The outcome is a pseudo-atomic model for 70% of the APC.

Within the APC, Doc1 aids substrate recognition⁶. Because this subunit reduces substrate dissociation, it allows the APC to be more processive — that is, to form longer ubiquitin chains before the substrate dissociates⁷. Most

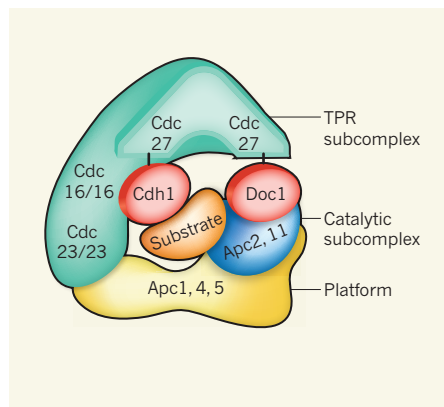


Figure 1 | Structure of the yeast APC¹⁻³. The platform — which consists of the Apc1, Apc4 and Apc5 subunits — connects the TPR and the catalytic subcomplexes. The catalytic subcomplex contains the cullin protein Apc2 and the RING protein Apc11. The TPR subcomplex contains two copies of each of the three TPR-containing subunits: Cdc23, 16 and 27. The terminal pair of Cdc27 subunits binds to the carboxy terminus of both another APC subunit, Doc1, and an activator protein such as Cdh1. These two proteins (red) both recognize the substrate.

APC substrates contain at least one of two APC-recognition motifs: the D-box and the KEN-box. Doc1 is thought to recognize the D-box⁸, whereas the APC activators Cdc20 and Cdh1 recognize both the D-box and the KEN-box.

Despite the suggestive data, however, evidence for a direct interaction between Doc1 and the D-box of substrate proteins has been lacking. By determining the EM structure of the APC with and without Doc1, the same group (da Fonseca *et al.*²; page 274) and another team (Buschhorn *et al.*³) map Doc1 to the central portion of the APC — between Cdc27 at the tip of the TPR subcomplex and the cullin Apc2 (Fig. 1). This positioning is consistent with previous data^{5,9} suggesting that Doc1 associates with both of these proteins, and places Doc1 adjacent to the activator Cdh1. Intriguingly, Doc1 seems to share some similarities with APC activators, including a carboxy-terminal motif that associates with the terminal TPR protein Cdc27^{5,9,10}. Because the APC has two copies of each of the TPR subunits, including Cdc27, da Fonseca *et al.* speculate that Doc1 and one of the activators each associate with separate Cdc27 subunits. Da Fonseca *et al.* also use nuclear magnetic resonance to confirm a direct interaction between Doc1 and D-box-containing peptides, although they could not map the binding surface.

Notably, both groups^{2,3} show that substrate arrival adds density between Doc1 and Cdh1. This occurs even when a peptide containing a single D-box is used as a substrate, hinting that Doc1 and Cdh1 together recognize a single site.

Earlier studies had shown that the APC exists

as a stable dimer that, like its monomeric form, can be purified. The dimeric form was found to be more processive than the monomeric form¹¹, leading to models in which the catalytic centres of each APC in the dimer are adjacent and affect substrate dissociation. The newly available EM structure of this dimer³ shows, however, that the two complexes are positioned back-to-back, with the catalytic centres facing away from one another. The tethered APCs are therefore unlikely to coordinate their efforts on one substrate. Nonetheless, in the dimer, the APC's structure differs slightly from that of the monomer, suggesting that dimerization alters APC conformation and so could affect its processivity.

Most ubiquitin ligases act as single-subunit enzymes that both recognize substrates and promote ubiquitination. Why the APC has evolved such a complex structure has therefore presented a puzzle. So far, defined functions have been established for only a few APC subunits. The new EM structures¹⁻³ are a good starting point for understanding whether the many subunits merely serve to bring the catalytic subcomplex close to the substrate or whether they have additional functions.

Other questions also remain. For one, exactly what is the defining characteristic of APC degradation signals — other than the D-box and the KEN-box — within substrates?

And do core APC subunits other than Doc1 recognize such signals? Does the APC dimer exist *in vivo*, and how does dimerization change the processivity of this complex? Answers to at least some of these questions will require further insight into the APC structure. ■

Ian Foe and David Toczyski are in the Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94158-9001, USA. e-mails: ian.foe@ucsf.edu; tooczyski@cc.ucsf.edu

- Schreiber, A. *et al.* *Nature* **470**, 227–232 (2011).
- da Fonseca, P. C. A. *et al.* *Nature* **470**, 274–278 (2011).
- Buschhorn, B. A. *et al.* *Nature Struct. Mol. Biol.* **18**, 6–13 (2011).
- Ohi, M. D. *et al.* *Mol. Cell* **28**, 871–885 (2007).
- Thornton, B. R. *et al.* *Genes Dev.* **20**, 449–460 (2006).
- Passmore, L. A. *et al.* *EMBO J.* **22**, 786–796 (2003).
- Carroll, C. W. & Morgan, D. O. *Nature Cell Biol.* **4**, 880–887 (2002).
- Carroll, C. W., Enquist-Newman, M. & Morgan, D. O. *Curr. Biol.* **15**, 11–18 (2005).
- Wendt, K. S. *et al.* *Nature Struct. Biol.* **8**, 784–788 (2001).
- Vodermaier, H. C., Gieffers, C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J.-M. *Curr. Biol.* **13**, 1459–1468 (2003).
- Passmore, L. A. *et al.* *Mol. Cell* **20**, 855–866 (2005).

ORGANIC CHEMISTRY

Metals are not the only catalysts

A long-standing problem in chemistry has been to find catalysts that allow molecules to distinguish between the two faces of reaction intermediates called carbocations. A way around the problem has been found. SEE LETTER P.245

MATTHEW GAUNT

The past 100 years have witnessed remarkable discoveries that have greatly advanced the synthetic use of alkenes — simple hydrocarbons that contain carbon-carbon double bonds (C=C bonds). As a result, alkenes are among the most frequently used building blocks for making organic molecules. Of particular note have been alkene reactions involving metal catalysts, a fruitful area of research leading to the award of three Nobel chemistry prizes¹ since 2001. But in this issue (page 245), Toste and colleagues² report a very different kind of catalyst for an alkene reaction. They have used a small organic molecule to catalyse the attachment of nitrogen-containing chemical groups to dienes, compounds that contain two C=C bonds. The resulting

products — cyclic molecules known as pyrrolidines — are widely found in biologically active molecules and could therefore be particularly useful in drug-discovery programmes.

Many alkene reactions generate chiral products that form as a mixture of two mirror-image isomers, known as enantiomers. Each enantiomer of a compound interacts with other chiral molecules in a specific way — sometimes with dramatic consequences, especially in biological systems. For example, the 'wrong' enantiomer of a drug molecule might be inactive, and at worst can cause severe side effects. The field of asymmetric synthesis, which aims to find reactions in which a single enantiomer of a product is formed, is therefore one of the liveliest and most important in chemistry.

Given the wide use of alkenes as molecular building blocks, discovering new ways to

transform them into more complex, enantiomerically pure molecules is one of the chief goals of asymmetric synthesis. Metal catalysts have been highly successful in this respect, and are often chemists' first port of call when developing such reactions. Toste and colleagues², however, report a new concept in the asymmetric catalysis of alkene reactions that represents a breakthrough for the field. Instead of a metal catalyst, they use a chiral form of a strong organic acid — a type of 'organocatalyst'^{3,4} — to make pyrrolidines and other cyclic molecules from dienes (Fig. 1).

Strong acids have long been known to catalyse reactions involving alkenes. When an alkene is mixed with a strong acid catalyst, a proton (H^+) is transferred from the acid to the alkene's C=C bond. The reactive species that results from this protonation is an intermediate called a carbocation (Fig. 1a). These are among the most versatile of reactive intermediates, not least because they can easily form chemical bonds to nucleophiles (molecules that are attracted to positive charges).

The protonation of alkenes is governed by a principle known as Markovnikov's rule⁵. This states that the positive charge in the resulting carbocation will be located on the carbon atom to which the fewest hydrogen atoms are attached, because that is where the charge becomes stabilized most effectively. For alkenes in which the C=C bond is at the end of a hydrocarbon chain, this often means that the charge will sit on an internal carbon of the intermediate, rather than at the end. The enantiomer of the product that forms when a nucleophile adds to the carbocation is then determined by which of the two faces of the intermediate is attacked. In most cases, the nucleophile can attack both faces equally well, and so a mixture of enantiomers forms. A long-standing challenge in asymmetric synthesis has been to make a single enantiomer of the product by directing the addition of a nucleophile to only one face of a carbocation. Solutions to this problem will have wide-ranging applications in chemical synthesis.

Toste and his team² have found a way around this synthetic conundrum. They used a special type of acid catalyst (a dithiophosphoric acid), which contains a chiral group that attaches covalently to the diene to form a reactive intermediate that mimics the reactivity of a carbocation (Fig. 1b). Once installed, the chiral group controls which side of the intermediate can be attacked by incoming nucleophiles. When the nucleophile is a sulphonamide (a nitrogen-containing group; Fig. 1c), a chiral pyrrolidine forms as the product, predominantly as one enantiomer; the chiral group is released as the product forms, regenerating the catalyst ready for another reaction.

The overall reaction is known as a hydroamination. Metal-catalysed hydroaminations are common, and in many cases yield products that have high enantioselectivities^{6,7}.

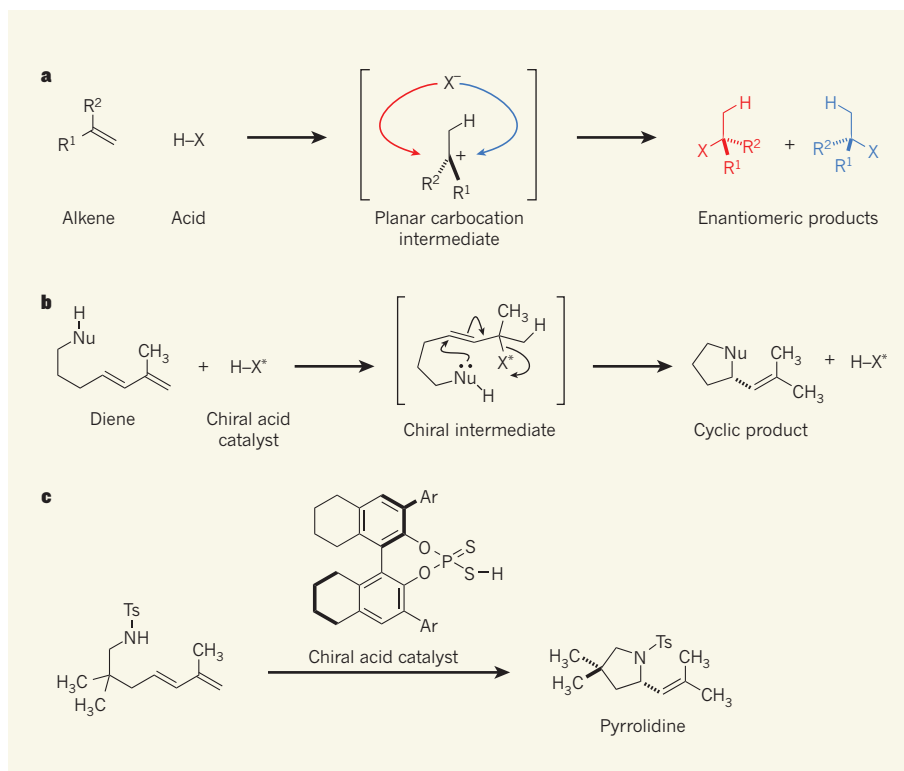


Figure 1 | Face selectivity in a reaction controlled by a chiral acid catalyst. **a**, The proton from strong acids can add to alkenes to yield planar intermediates known as carbocations. The counterion of the acid (X^-) then attacks the intermediate; a different enantiomer of the final product forms depending on which face of the carbocation the counterion attacks. Because the counterion can approach either face equally easily, a mixture of enantiomers is produced. R^1 and R^2 are hydrocarbon groups. **b**, Toste and colleagues² report a reaction in which a diene is protonated by an acid that has a chiral counterion. The counterion forms a covalent bond to the substrate, generating a chiral intermediate that mimics the reactivity of the carbocations in **a**. A nucleophile (Nu) in the substrate then attacks the carbon–carbon double bond in the intermediate from one side only, because the chiral X^* group blocks the other face of the molecule. A cyclic product forms, and the acid catalyst is regenerated. Curly arrows indicate electron movement; the pair of dots on the nucleophile represents a pair of electrons. **c**, Using this approach, Toste *et al.* have made pyrrolidine heterocycles essentially as single enantiomers. Ar represents aromatic groups; Ts is an arylsulphonyl group, $SO_2C_6H_4CH_3$.

Acid-catalysed hydroaminations are also well known^{8–10}, but Toste and colleagues' reaction is the first highly enantioselective variant.

The authors demonstrate² that their reaction works for a range of sulphonamides and dienes. At present, the reaction seems to work best when making five-membered rings, for which excellent enantioselectivities are obtained. But it also has the potential to work well for larger ring sizes, although further development will be needed to make these important compounds.

Toste and colleagues also outline an exciting preliminary result showing that the nucleophile need not be a nitrogen atom. When an aromatic molecule, indole, is incorporated into the starting material in place of the sulphonamide group, a carbon atom in the indole acts as a nucleophile instead (see Fig. 2e on page 248). The resulting cyclic products contain 'privileged' structures — molecular scaffolds that are known to be especially active at a wide variety of drug targets. Perhaps more importantly, however, the results with indole suggest that

chiral-acid-catalysed reactions could be applicable to a wide range of nucleophiles. If so, the authors have discovered a general concept in catalysis that will have a major impact on how chiral molecules are made. ■

Matthew Gaunt is in the Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK.
e-mail: mjj32@cam.ac.uk

1. http://nobelprize.org/nobel_prizes/chemistry/laureates
2. Shapiro, N. D., Rauniyar, V., Hamilton, G. L., Wu, J. & Toste, F. D. *Nature* **470**, 245–249 (2011).
3. MacMillan, D. W. C. *Nature* **455**, 304–308 (2008).
4. Terada, M. *Chem. Commun.* 4097–4112 (2008).
5. Markownikoff, W. *Justus Liebigs Ann. Chem.* [now *Eur. J. Org. Chem.*] **153**, 228–259 (1870).
6. Aillaud, I., Collin, J., Hannedouche, J. & Schulz, E. *Dalton Trans.* 5105–5118 (2007).
7. Hultsch, K. C. *Adv. Synth. Catal.* **347**, 367–391 (2005).
8. Schlummer, B. & Hartwig, J. F. *Org. Lett.* **4**, 1471–1474 (2002).
9. Li, Z. *et al. Org. Lett.* **8**, 4175–4178 (2006).
10. Ackermann, L. & Althammer, A. *Synlett* 995–998 (2008).

Initial impact of the sequencing of the human genome

Eric S. Lander¹

The sequence of the human genome has dramatically accelerated biomedical research. Here I explore its impact, in the decade since its publication, on our understanding of the biological functions encoded in the genome, on the biological basis of inherited diseases and cancer, and on the evolution and history of the human species. I also discuss the road ahead in fulfilling the promise of genomics for medicine.

On 15 February 2001, a decade ago this week, *Nature* published a 62-page paper entitled 'Initial sequencing and analysis of the human genome', reporting a first global look at the contents of the human genetic code. The paper¹ marked a milestone in the international Human Genome Project (HGP), a discovery programme conceived in the mid-1980s and launched in 1990. The same week, *Science* published a paper² from the company Celera Genomics, reporting a draft human sequence based on their own prodigious data, as well as data from the public HGP.

The human genome has had a certain tendency to incite passion and excess: from early jeremiads that the HGP would strangle research by consuming the NIH budget (it never rose to more than 1.5%); to frenzied coverage of a late-breaking genome race between public and private protagonists; to a White House announcement of the draft human sequence in June 2000, 8 months before scientific papers had actually been written, peer-reviewed and published; to breathless promises from Wall Street and the press about the imminence of genetic 'crystal balls' and genome-based panaceas; to a front-page news story on the tenth anniversary of the announcement that chided genome scientists for not yet having cured most diseases.

The goal of this review is to step back and assess the fruits of the HGP from a scientific standpoint, addressing three questions: what have we learned about the human genome itself over the past decade? How has the human sequence propelled our understanding of human biology, medicine, evolution and history? What is the road ahead?

The past decade has shown the power of genomic maps and catalogues for biomedical research. By providing a comprehensive scaffold, the human sequence has made it possible for scientists to assemble often fragmentary information into landscapes of biological structure and function: maps of evolutionary conservation, gene transcription, chromatin structure, methylation patterns, genetic variation, recombinational distance, linkage disequilibrium, association to inherited diseases, genetic alterations in cancer, selective sweeps during human history and three-dimensional organization in the nucleus. By providing a framework to cross-reference information across species, it has connected the biology of model systems to the physiology of the human. Furthermore, by providing comprehensive catalogues of genomic information, it has enabled genes and proteins to be recognized based on unique 'tags'—allowing, for example, RNA transcripts to be assayed with arrays of oligonucleotide probes and proteins by detection of short peptide fragments in a mass spectrometer. In turn, these measurements have been used to construct 'cellular signatures' characteristic of specific cell types, states and responses, and catalogues of the contents of organelles such as the mitochondria.

The intensity of interest can be seen in the 2.5 million queries per week on the major genome data servers and in the flowering of a rich field of computational biology.

The greatest impact of genomics has been the ability to investigate biological phenomena in a comprehensive, unbiased, hypothesis-free manner. In basic biology, it has reshaped our view of genome physiology, including the roles of protein-coding genes, non-coding RNAs and regulatory sequences.

In medicine, genomics has provided the first systematic approaches to discover the genes and cellular pathways underlying disease. Whereas candidate gene studies yielded slow progress, comprehensive approaches have resulted in the identification of ~2,850 genes underlying rare Mendelian diseases, ~1,100 loci affecting common polygenic disorders and ~150 new recurrent targets of somatic mutation in cancer. These discoveries are propelling research throughout academia and industry.

The following sections contain only a small number of citations due to space limitations; a more extensive bibliography tied to each section can be found as Supplementary Information.

Genome sequencing

The view from 2000

Genome sequencing was a daunting task in late 2000. The catalogue of organisms with published genome sequences was small: thirty-eight bacteria, one fungus (*Saccharomyces cerevisiae*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*) and one plant (*Arabidopsis thaliana*), all with relatively small and simple genomes.

The human genome was much more challenging, being roughly an order of magnitude larger than the total of all previous genomes and filled with repetitive sequences. The human draft sequence reported in 2001, although a landmark, was still highly imperfect. It covered only ~90% of the euchromatic genome, was interrupted by ~250,000 gaps, and contained many errors in the nucleotide sequence¹.

Finishing the human

After the draft sequence, the HGP consortium quickly turned to producing a high-quality reference sequence, through lapidary attention to individual clones spanning the genome. In 2004, it published a near-complete sequence that was a vast improvement³: it contained ~99.7% of the euchromatic genome, was interrupted by only ~300 gaps, and harboured only one nucleotide error per 100,000 bases. The gaps consisted of ~28 megabases (Mb) of euchromatic sequence, mostly involving repetitive regions that could not be reliably cloned or assembled, and ~200 Mb of heterochromatic sequence, including the large centromeres and the short arms of acrocentric chromosomes.

¹Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.

This 'finished' sequence allowed accurate inference of gene structure and detection of polymorphisms and mutation across the genome. Importantly, the clone-based approach also provided access to regions with recent segmental duplications, which are poorly represented in whole-genome shotgun sequencing. These segmental-duplication-rich regions have extraordinarily high rates of copy-number variation within and between species, and they mediate large-scale chromosomal rearrangement. In addition, they are nurseries for rapidly evolving gene families under positive selection, and they account for a disproportionate share of disease burden, particularly for neurological and neuropsychiatric disorders.

Expanding the bestiary

Additional vertebrate genomes followed rapidly, to help interpret the human genome through comparative analysis and to enable experimental studies in model systems. Key genomes included mouse⁴, dog⁵, rat⁶, chimpanzee⁷ and cow, as well as a marsupial⁸, monotreme and bird. Remarkably, partial genome sequences have even been obtained from several extinct species, notably the woolly mammoth and our closest relative, Neanderthal⁹. The current catalogue includes ~250 eukaryotes (totalling ~120 gigabases (Gb)) and ~4,000 bacteria and viruses (~5 Gb). Sequencing has extended to microbial communities, including samples from the mid-ocean and environmental remediation sites and human samples from gut and skin^{10,11}.

Massively parallel sequencing

The HGP used essentially the same sequencing method introduced by Sanger in 1977: electrophoretic separation of mixtures of randomly terminated extension products, although dramatically improved with fluorescently labelled terminators and automated laser detectors. The past half-decade has seen a tectonic shift in sequencing technology, based on *in situ* sequencing in which two-dimensional optical imaging is used to monitor sequential addition of nucleotides to spatially arrayed DNA templates. Whereas electrophoretic methods could deploy ~10² parallel channels, optical imaging can now follow ~10⁹ templates. The per-base cost of DNA sequencing has plummeted by ~100,000-fold over the past decade, far outpacing Moore's law of technological advance in the semiconductor industry. The current generation of machines can read ~250 billion bases in a week, compared to ~25,000 in 1990 and ~5 million in 2000.

A drawback of the new technology is that the sequence reads are much shorter than the ~700 bases routinely provided by electrophoretic methods. Because it is challenging to assemble a genome sequence *de novo* from such short reads, most applications have focused on placing reads onto the scaffold of an existing genome sequence to count their density or to look for differences from a reference sequence.

Applications

An early application of massively parallel sequencing was to create 'epigenomic maps', showing the locations of specific DNA modifications, chromatin modifications and protein-binding events across the human genome. Chromatin modification and protein binding can be mapped by chromatin immunoprecipitation-sequencing (ChIP-Seq)^{12,13}, and the sites of DNA methylation can be found by sequencing DNA in which the methylated cytosines have been chemically modified (Methyl-Seq)¹⁴.

As the technology has improved, the focus has turned to re-sequencing human samples to study inherited variation or somatic mutations. One can re-sequence the whole genome¹⁵ to varying degrees of coverage or use hybridization-capture techniques¹⁶ to re-sequence a targeted subset, such as the protein-coding sequences (referred to as the 'exome').

Sequencing is also being extensively applied to RNA transcripts (RNA-Seq), to count their abundance, identify novel splice forms or spot mutations¹⁷. A harder challenge is reconstructing a transcriptome *de novo*, but good algorithms have recently been developed^{18,19}.

The hardest challenge is *de novo* assembly of entire genomes, but even here there has been recent progress in achieving long-range connectivity.

For the human, initial efforts yielded scaffolds of modest size (~500 kb), and recent algorithms²⁰ have approached the typical range for capillary-based sequencing (11.5-Mb scaffolds, containing ~90% of the genome). Encouraged by this progress, a scientific consortium has begun laying a plan to sequence 10,000 vertebrate genomes—roughly one from every genus.

The road ahead

The ultimate goal is for sequencing to become so simple and inexpensive that it can be routinely deployed as a general-purpose tool throughout biomedicine. Medical applications will eventually include characterizing patients' germline genomes (to detect strongly predictive mutations for presymptomatic counselling where treatments exist, to search for causes of disease of unknown aetiology, and to detect heterozygous carriers for prenatal counselling); cancer genomes (by identifying somatic mutations to compare tumour and normal DNA); immune repertoires (by reading the patterns of B-cell and T-cell receptors to infer disease exposures and monitor responses to vaccines); and microbiomes (by associating patterns of microbial communities with diseases processes). Research applications will include characterizing genomes, epigenomes and transcriptomes of humans and other species, as well as using sequencing as a proxy to probe diverse molecular interactions.

To fulfil this potential, the cost of whole-genome sequencing will need eventually to approach a few hundred US dollars. With new approaches under development and market-based competition, these goals may be feasible within the next decade.

Genome anatomy and physiology

The view from 2000

Our knowledge of the contents of the human genome in 2000 was surprisingly limited. The estimated count of protein-coding genes fluctuated wildly. Protein-coding information was thought to far outweigh regulatory information, with the latter consisting largely of a few promoters and enhancers per gene. The role of non-coding RNAs was largely confined to a few classical cellular processes. And, the transposable elements were largely regarded as genomic parasites.

A decade later, we know that all of these statements are false. The genome is far more complex than imagined, but ultimately more comprehensible because the new insights help us to imagine how the genome could evolve and function.

Protein-coding genes

Since the early 1970s, the total number of genes (the vast majority assumed to be protein-coding) had been variously estimated at anywhere from ~35,000 to well over 100,000, based on genetic load arguments, hybridization experiments, the average size of genes, the number of CpG islands and shotgun sequencing of expressed sequence tags. The HGP paper suggested a total of 30,000–40,000 protein-coding genes, but the estimate involved considerable guesswork owing to the imperfections of the draft sequence and the inherent difficulty of gene identification.

Today, the human genome is known to contain only ~21,000 distinct protein-coding genes²¹. Generating a reliable gene catalogue required eliminating the many open reading frames (ORFs) that occur at random in transcripts, while retaining those that encode bona fide proteins. The key insight was to identify those ORFs with the evolutionary signatures of bona fide protein-coding genes (such as amino-acid-preserving substitutions and reading-frame-preserving deletions) and prove that most ORFs without such conservation are not newly arising protein-coding genes. Recent RNA-Seq projects have confirmed the gene catalogue, while illuminating alternative splicing, which seems to occur at >90% of protein-coding genes and results in many more proteins than genes.

The proteome is now known to be similar across placental mammals, with about two-thirds of protein-coding genes having 1:1 orthologues across species and most of the rest belonging to gene families that undergo regular duplication and divergence—the invention of fundamentally new proteins is rare.

Conserved non-coding elements

The most surprising discovery about the human genome was that the majority of the functional sequence does not encode proteins. These features had been missed by decades of molecular biology, because scientists had no clue where to look.

Comparison of the human and mouse genomes showed a substantial excess of conserved sequence, relative to the neutral rate in ancestral repeat elements⁴. The excess implied that at least 6% of the human genome was under purifying selection over the past 100 million years and thus biologically functional. Protein-coding sequences, which comprise only ~1.5% of the genome, are thus dwarfed by functional conserved non-coding elements (CNEs). Subsequent comparison with the rat and dog genomes confirmed these findings^{5,6}.

Although the initial analysis provided a bulk estimate of the amount of conserved sequence, it could only pinpoint the most highly conserved elements. Among them are nearly 500 ultraconserved elements (200 bases or more perfectly conserved across human, mouse and rat), most of which neither overlap protein-coding exons nor show evidence of being transcribed²². On the basis of statistical measures of constraint, tens of thousands of additional highly conserved non-coding elements (HCNEs) were identified^{5,23}. In many cases the evolutionary origins of these HCNEs could be traced back to the common ancestor of human and fish. HCNEs preferentially reside in the gene deserts that often flank genes with key functions in embryonic development^{5,22,23}. Large-scale screens of these sequences in transgenic mice revealed that they are highly enriched in tissue-specific transcriptional enhancers active during embryonic development²⁴, revealing a stunning complexity of the gene regulatory architecture active in early development²⁴ (Fig. 1).

Sequencing additional genomes has gradually increased our power to pinpoint the less stringently conserved CNEs. Recent comparison with 29 mammalian genomes has identified millions of additional conserved elements, comprising about two-thirds of the total conserved sequence.

Evolutionary analyses of CNEs have also enabled the discovery of distinct types of functional elements, including regulatory motifs present in the promoters and untranslated regions of co-regulated genes, insulators that constrain domains of gene expression, and families of conserved secondary structures in RNAs. Nonetheless, the function of most CNEs remains to be discovered.

Nature of evolutionary innovation

Although some CNEs show deep conservation across vertebrate evolution, most evolve and turn over at a faster pace than protein-coding sequences. At least 20% of the CNEs conserved among placental mammals are absent in marsupial mammals, compared to only ~1% for protein-coding sequences⁸. These elements arose in the period between our common ancestor with marsupial mammals (~180 million years (Myr) ago) and our common ancestor with placental mammals (~90 Myr ago), or else were ancestral elements lost in the marsupial lineage. The proportion of CNEs having detectable conservation with birds is much lower (~30% detectable, ~310 Myr ago) and with fish is near zero

(~450 Myr ago). The more rapid change of CNEs provides experimental support for the notion²⁵ that evolution of species depends more on innovation in regulatory sequences than changes in proteins.

Extrapolation from the marsupial-placental comparison indicates that at least ~20% of functional non-coding elements in human should be absent in mouse. Interestingly, ChIP-Seq studies report even greater differences in the localization of transcription factor binding sites between mammals. However, physical binding may not imply biological function: the evolutionary and biochemical data still need to be reconciled.

Transposons as drivers of evolutionary innovation

The HGP paper included a detailed analysis of transposon-derived sequences, but largely viewed transposons as a burden on the genome.

Comparative genomics, however, began to change this picture. The first hint was a handful of families of CNEs that had clearly been derived from transposons²⁶. Comparison of placental and marsupial genomes then revealed that at least 15% of the CNEs that arose during the period from 180 Myr ago to 90 Myr ago were derived from transposon sequences⁸; the true total is likely to be considerably larger, because the flanking transposon-derived sequences will have degenerated in many cases. In retrospect, the advantage seems obvious. First, most transposons contain sequences that interact with the host transcriptional machinery, and therefore provide a useful substrate for evolution of novel regulatory elements. Second, a regulatory control that evolved at one locus could give rise to coordinated regulation across the genome by being picked up by a transposon, scattered around the genome and retained in advantageous locations. Over evolutionary timescales, transposons may earn their keep.

Small non-coding RNAs

The HGP paper analysed all known classes of functional human non-protein-coding RNAs (ncRNAs), which consisted largely of those supporting protein translation (ribosomal, transfer and small nucleolar RNAs) and transcript splicing (small nuclear RNAs).

In late 2000, vertebrates were found to harbour an important new type of ncRNA first discovered in *C. elegans*. Called microRNAs (miRNAs), these products bind target mRNAs and decrease their stability²⁷. Today, the human genome is known to encode ~100 evolutionarily conserved families of miRNAs. Genomic analysis proved critical in identifying the target mRNAs: evolutionarily conserved 7-base sequences in 3' untranslated regions complementary to bases 2–8 of conserved miRNAs²⁸. A typical conserved miRNA has ~200 target mRNAs with conserved binding sites. A few dozen miRNAs have been shown to have key regulatory roles, such as in cancer and development. Many of the others may help to fine-tune gene expression, although some may be too subtle to have detectable phenotypes in laboratory experiments. Recently, a new class of small RNAs, called PIWI-interacting RNAs, has been discovered that functions through a similar molecular machinery—they act to silence transposons in the germline.

Ubiquitous transcription

In 2000, transcription was thought to be largely confined to regions containing protein-coding genes. Only a handful of non-classical large functional ncRNAs was known, such as telomerase RNA, 7SL signal recognition RNA, Xist and H19, and these were regarded as quirky exceptions. Pioneering studies of the human sequence soon began to provide hints that additional large RNA molecules might exist. Hybridization of RNA to microarrays of genomic sequence suggested that more than 10% of the genome was represented in mature transcripts, with most lying outside protein-coding exons²⁹, and random cDNA sequencing turned up many transcripts that could not be linked to protein-coding genes³⁰ (Fig. 2). With increasingly sensitive assays, it was concluded by 2007 that virtually every nucleotide in the euchromatic genome was likely to be represented in primary (unspliced) transcripts in at least some cell type at some time³¹. Many of these transcripts, however, have extremely low expression levels and show little evolutionary

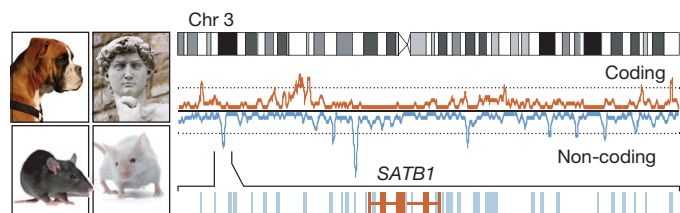


Figure 1 | Evolutionary conservation maps. Comparison among the human, mouse, rat and dog genomes helps identify functional elements in the genome. The figure shows the density of protein-coding sequences (red) and the most highly conserved non-coding sequences (blue) along chromosome 3. Highly conserved non-coding sequences are enriched in gene-poor regions, each of which contained a gene involved in early development (such as *SATB1*, shown). Images courtesy of iStock Photo.

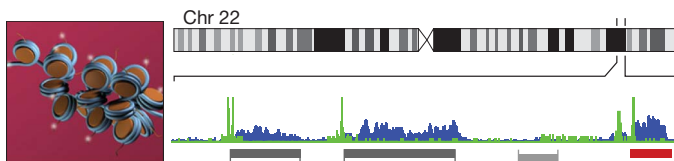


Figure 2 | Chromatin state maps. The genomic sites of chromatin modifications or protein binding can be mapped, using chromatin immunoprecipitation (ChIP) and massively parallel sequencing. The figure highlights chromatin marks associated with the active promoters (green) and actively transcribed regions (blue), in a region on chromosome 22. The four features shown correspond to two active protein-coding (dark grey), one inactive protein-coding (light grey) and one long intergenic non-coding RNA (maroon). Image courtesy of B. Wong (ClearScience).

conservation; these may represent ‘transcriptional noise’ (that is, reproducible, tissue-specific transcription from loci with randomly occurring weak regulatory signals). Exactly how much of the ubiquitous transcription is biologically functional remains controversial.

Large intergenic non-coding RNAs

Epigenomic maps facilitated the discovery of a large class of thousands of genes encoding evolutionarily conserved (and thus clearly functional) transcripts, now called large intergenic non-coding RNAs (lincRNAs)³². The genes were pinpointed because they carry the distinctive chromatin patterns of actively transcribed genes but lack any apparent protein-coding capacity³². On the basis of their expression patterns, they have diverse roles in processes such as cell-cycle regulation, immune responses, brain processes and gametogenesis. A substantial fraction binds chromatin-modifying proteins and may modulate gene expression, for example, in the HOX complex³³ and in the p53-response pathway. Although their mechanism of action remains to be elucidated, lincRNAs may act analogously to telomerase RNA by serving as ‘flexible scaffolds’³⁴ that bring together protein complexes to elicit a specific function.

lincRNAs are not the end of the ncRNA story. RNA-Seq studies have begun to define catalogues of antisense RNAs that overlap protein-coding genes¹⁸. Unlike lincRNAs, these transcripts show little evolutionary conservation (beyond the coding region)¹⁸ and may function by base-pairing with the overlapping transcript or simply by causing chromatin changes through the act of transcription.

Epigenomic maps

Recognizing the distinctive functional domains in the genome of a cell is a key challenge, both for genome scientists and for the cell itself. With thousands of genome-wide epigenomic maps, it is now clear that functionally active domains are associated with specific patterns of epigenomic marks^{12,13,31,35} (Fig. 2). For example, active promoters show DNase hypersensitivity, histone acetylation and histone 3 lysine 4 trimethylation; transcribed regions are marked by histone 3 lysine 4 trimethylation; and enhancers show binding of the p300 acetyltransferase. Other features are seen at exons, insulators and imprinting control regions. The binding sites of transcription factors can also be read out, given an antibody with adequate specificity.

Moreover, it is possible to study dynamic behaviour and developmental potential by comparing epigenomic maps from related cellular states. For example, bivalent chromatin domains (both histone 3 lysine 4 trimethylation and histone 3 lysine 27 trimethylation) mark genes that are poised to play key parts in subsequent lineage decisions³⁶. Epigenomic maps can also reveal genes that serve as obstacles to cellular reprogramming, and DNA methylation maps are helping identify aberrant functions in cancer³⁷. Ultimately, hundreds of thousands of epigenomic marks will be layered atop the genome sequence to provide an exquisite description of genomic physiology in a cell type.

Three-dimensional structure of the genome

Whereas general features of chromosomal packaging had been worked out through classical techniques such as X-ray diffraction, little was

known about *in vivo* physical contacts between genomic loci more than a few kilobases apart.

The (one-dimensional) genome sequence enabled technologies for mapping the genome in three dimensions. Chromosome conformation capture (3C) could test whether two loci are nearby in the nucleus, based on proximity-based ligation followed by locus-specific polymerase chain reaction³⁸. It revealed, for example, that β -globin’s locus control region forms an ‘active hub’ involving physical contact between genomic elements separated by 100 kb or more.

New approaches, such as a method called Hi-C, extend 3C to examine all physical contacts in an unbiased genome-wide fashion³⁹. It has revealed that the genome is organized into two compartments, corresponding to open and closed chromatin, and, at megabase scale, exhibits folding properties consistent with an elegant structure called a fractal globule.

The road ahead

The ultimate goal is to understand all of the functional elements encoded in the human genome. Over the next decade, there are two key challenges. The first will be to create comprehensive catalogues across a wide range of cell types and conditions of (1) all protein-coding and non-coding transcripts; (2) all long-range genomic interactions; (3) all epigenomic modifications; and (4) all interactions among proteins, RNA and DNA. Some efforts, such as the ENCODE and Epigenomics Roadmap projects, are already underway³¹. Among other things, these catalogues should help researchers to infer the biological functions of elements; for example, by correlating the chromatin states of enhancers with the transcriptional activity of nearby genes across cell types and conditions. These goals should be feasible with massively parallel sequencing and assay miniaturization, although they will require powerful ways to purify specific cell types *in vivo*, and the fourth goal will require a concerted effort to generate specific affinity reagents that recognize the thousands of proteins that interact with nucleic acids.

The second and harder challenge is to learn the underlying grammar of regulatory interactions; that is, how genomic elements such as promoters and enhancers act as ‘processors’ that integrate diverse signals. Large-scale observational data will not be enough. We will need to engage in large-scale design, using synthetic biology to create, test and iteratively refine regulatory elements. Only when we can write regulatory elements *de novo* will we truly understand how they work.

Genomic variation

The view from 2000

Since the early 1980s, humans were known to carry a heterozygous site roughly every 1,300 bases. Genetic maps containing a few thousand markers, adequate for rudimentary linkage mapping of Mendelian diseases, were constructed in the late 1980s and early 1990s. Systematic methods to discover and catalogue single nucleotide polymorphisms (SNPs) were developed in the late 1990s and resulted in the report of 1.42 million genetic variants in a companion to the HGP paper⁴⁰. Still, the list was far from complete. Moreover, there was no way to actually assay the genotypes of these SNPs in human samples.

Today, the vast majority of human variants with frequency >5% have been discovered and 95% of heterozygous SNPs in an individual are represented in current databases. Moreover, geneticists can readily assay millions of SNPs in an individual.

Linkage disequilibrium, HapMaps and SNP chips

Two critical advances propelled progress in the study of genomic variation: one conceptual and one technical. The first was the discovery of the haplotype structure of the human genome⁴¹; that is, that genetic variants in a region are tightly correlated in structures called haplotypes, reflecting linkage disequilibrium and separated by hotspots of recombination. Linkage disequilibrium was a classical concept, but its genome-wide structure had never been characterized in any organism. Humans turned out to have a surprisingly simple structure, reflecting recent expansion from a

small founding population. Tight correlations seen in a few dozen regions⁴¹ implied that a limited set of ~500,000–1,000,000 SNPs could capture ~90% of the genetic variation in the population. The International Haplotype Map (HapMap) Project soon defined these patterns across the entire genome, by genotyping ~3 million SNPs⁴². The second advance was the development of genotyping arrays (often called SNP chips), which can now assay up to ~2 million variants simultaneously.

Copy-number polymorphisms

Large-scale genomic aberrations (deletions and duplications) were long known to occur in cancers and congenital disorders. Biologists using DNA microarrays to study these events made a surprising observation: even normal, apparently healthy individuals showed copy-number polymorphism (CNP) in many genomic segments⁴³. A typical person carries ~100 heterozygous CNPs covering ~3 Mb; the figure is vastly lower than initial estimates but still considerable⁴⁴. Most are ancient variants that are tightly correlated with SNPs, which has enabled the association of CNVs with phenotypes using proxy SNPs. An intriguing minority of CNVs, however, seems to arise from recent and *de novo* mutations and may have important roles in psychiatric disorders^{45–47}.

The road ahead

The ultimate goal is to create a reference catalogue of all genetic variants common enough to be encountered recurrently in populations, so that they can be examined for association with phenotypes and interpreted in clinical settings. Efforts towards this goal are already well underway. The 1000 Genomes Project⁴⁸ (which plans to study many more than one thousand genomes) aims to find essentially all variants with frequency >1% across the genome and >0.1% in protein-coding regions.

Medicine: Mendelian and chromosomal disorders

The view from 2000

At the time when the HGP was launched, fewer than 100 disease genes had been identified, because finding them largely relied on guesswork about the underlying biochemistry. Genetic linkage mapping in affected families offered a general solution in principle, but it was slow and tedious. With the genetic and physical maps created in the first stages of the HGP, the list of identified disease genes quickly began to grow. With the human sequence, it has exploded. A decade after the HGP, more than 2,850 Mendelian disease genes have been identified.

Mendelian diseases

Given enough affected families, one can genetically map a monogenic disease to a chromosomal region and compare patient and reference sequences to search for the causative gene and mutation. With massively parallel sequencing, the simplest way to interrogate a region is often by whole-exome or whole-genome re-sequencing. Increasingly, investigators are attempting to eliminate the step of genetic mapping. The task is not entirely straightforward: even when all common variants can be filtered out based on the fruits of the 1000 Genomes Project, a typical person will still have ~150 rare coding variants affecting ~1% of their genes (as well as 100-fold more rare non-coding variants). For recessive diseases caused by protein-coding mutations, it may just be possible to discover a disease gene based on a single patient by looking for two mutations in the same gene. In general, though, pinpointing the right gene will require accumulating evidence from multiple patients.

Clinical applications

DNA sequencing is being increasingly used in the clinic, including applying whole-exome sequencing to assign patients with an unclear diagnosis to a known disease⁴⁹. Genomics is also becoming a routine tool in cytogenetics laboratories, where DNA microarrays have greatly increased the sensitivity to detect clinically significant chromosomal imbalances, advancing diagnostic evaluation of children with idiopathic developmental delay, major intellectual disability, autism and birth defects.

The road ahead

New sequencing technologies should propel tremendous progress in elucidating the ~1,800 uncloned disorders in the current catalogue, and in recognizing undescribed Mendelian disorders in patients with unexplained congenital conditions. By comparing patients to parents using highly accurate sequencing, it will even be possible to spot newly arising mutations responsible for dominant lethal disorders. As whole-genome sequencing costs fall, it may become routinely used by couples before conception, and by paediatricians to explain idiopathic conditions in children.

Because Mendelian recessive disorders often involve the complete absence of a protein, they are typically difficult to treat, except where enzyme replacement is feasible. Genomics has produced some remarkable recent exceptions, such as the mechanism-based recognition that Marfan syndrome might be treated with an existing inhibitor of TGF- β ⁵⁰. A critical challenge will be to find systematic ways to treat Mendelian disorders by gene-based therapies, or by developing small-molecule therapeutics.

Medicine: common diseases and traits

The view from 2000

In contrast to rare Mendelian diseases, extensive family-based linkage analysis in the 1990s was largely unsuccessful in uncovering the basis of common diseases that afflict most of the population. These diseases are polygenic, and there were no systematic methods for identifying underlying genes. As of 2000, only about a dozen genetic variants (outside the HLA locus) had been reproducibly associated with common disorders.

A decade later, more than 1,100 loci affecting more than 165 diseases and traits have been associated with common traits and diseases, nearly all since 2007.

Common disease

To study common diseases, geneticists conceived principles for genetic mapping using populations rather than families. A first systematic example was genome-wide association study (GWAS), which involves testing a comprehensive catalogue of common genetic variants in cases and controls from a population to find those variants associated with a disease^{51,52}. Rare Mendelian diseases are almost always caused by a spectrum of rare mutations, because selection acts strongly against these alleles. By contrast, the 'common disease–common variant' (CD/CV) hypothesis⁵³ posited that common genetic variants (polymorphisms, classically defined as allele frequency >1%) could have a role in the aetiology of common diseases. By testing all common variants, one could pinpoint key genes and shed light on underlying mechanisms.

The CD/CV hypothesis rested on the following premise: because the vast majority (~99%) of genetic variance in the population is due to common variants, the susceptibility alleles for a trait will include many common variants except if the alleles have had a large deleterious effect on reproductive fitness over long periods. For common diseases or traits, many susceptibility alleles may have been only mildly deleterious, neutral or even advantageous. Examples may include diseases of late onset, diseases resulting from recent changes in living conditions such as diabetes and heart disease, morphological traits, and alleles with pleiotropic effects that result in balancing selection. Notably, humans are a favourable case for genetic mapping by association studies, because the small historical population size means that the force of selection is weaker and the allelic spectrum is simpler⁵³.

With the development of catalogues of common variants, haplotype maps, genotyping arrays and rigorous statistical methods⁵⁴, the CD/CV hypothesis was finally put to the test beginning in late 2006; it has been richly confirmed by an explosion of discoveries.

Revealing disease pathways

Three key results have emerged from these studies: (1) most traits can be influenced by a large number of loci; (2) the vast majority of the common variants at these loci have a moderate effect, increasing risk by 10–50% (similar to effects of many environmental risk factors); and (3) the loci

include most of the genes found by linkage analysis, but reveal many more genes not previously implicated. Some early commentators argued that such discoveries were not useful for understanding disease, because loci with moderate effects were too hard to study and could not have important therapeutic consequences. The results, however, have proved otherwise.

By discovering large collections of genes that can modulate a phenotype, GWAS has begun to reveal underlying cellular pathways and, in some cases, already pointed to new therapeutic approaches. In effect, GWAS is the human analogue to mutagenesis experiments in animal models: they provide a systematic, unbiased way to identify genes and pathways underlying a biological process to allow subsequent physiological studies. A recent study of the genetic control of lipid levels illustrates many of the points (Box 1). Some other important examples are given in the following paragraphs.

Adult macular degeneration. GWAS uncovered the aetiology of this leading cause of blindness affecting millions of elderly patients, by finding strong associations with five loci with common variants of large effect, including multiple genes in the alternative complement system. The results pointed to a failure to inhibit specific inflammatory responses, spurring new therapeutic approaches.

BOX 1

Genetics of lipids

A recent GWAS of plasma lipid levels, a major risk factor for myocardial infarction, demonstrates the power of the approach. A study of >100,000 individuals of European ancestry identified 95 loci associated to at least one of three major lipids: low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides⁶⁶. Moreover, most also showed association in African and Asian cohorts, indicating their generality.

Although the variants have only moderate effects, their combined impact can be considerable. Together, the loci explain ~25% of the genetic variance for LDL and HDL levels. Furthermore, the top quartile of individuals with triglyceride-associated variants has a 44-fold higher risk of hypertriglyceridaemia than the bottom quartile.

Notably, the 95 loci with common variants include nearly all of the 18 genes previously implicated in rare Mendelian lipid disorders, indicating that GWAS will often help to pinpoint genes harbouring rare variants. Of note, at loci where both common and rare variants were studied, the former explain much more of the heritability than the latter.

The study underscores that loci with only moderate effects in GWAS may have major therapeutic implications. The *HMGCR* locus has a common variant at 40% frequency that changes LDL by a modest 2.8 mg dl⁻¹ and no known rare mutations of large effect, presumably because they would be lethal. Yet, the encoded protein is the target of statins, drugs taken by tens of millions of patients that can significantly reduce both LDL levels and myocardial infarction risk.

A number of the new loci identified have already been confirmed to affect lipid biology, through rapid transgenic animal studies and human clinical studies. The sortilin gene (*SORT1*), for example, contains a common variant that creates a novel transcription-factor-binding site that alters hepatic expression in humans, and transgenic studies in mouse show that it alters plasma LDL levels¹⁰⁰. The gene thus defines a previously unknown regulatory pathway for LDL.

Separate studies have identified a pair of common nonsense variants in the *PCSK9* gene in African-Americans (2.6% combined frequency) that markedly reduces both LDL levels and risk of myocardial infarction⁶⁹. The finding that human homozygotes for this presumably null allele are healthy indicates that a drug against the encoded protein should be safe and effective, provoking considerable activity by pharmaceutical companies.

Crohn's disease. Studies have so far identified 71 risk loci⁵⁵ for this severe inflammatory bowel disease (Fig. 3). The genes have together revealed previously unknown roles for such processes as innate immunity, autophagy and interleukin-23 receptor signalling. Specific mutations identified in patients have been confirmed to be pathogenic in cellular and animal models, providing new strategies for therapeutic development.

Control of fetal haemoglobin. Fetal haemoglobin (HbF) levels vary among individuals and higher levels can ameliorate symptoms of sickle cell anaemia and β -thalassaemia, but hopes for treating these diseases by increasing HbF had been stymied by ignorance of the mechanism that downregulates HbF expression. GWAS revealed three loci that modulate erythroid development, which together explain >25% of the genetic variance in HbF levels and are associated with reduced severity in sickle cell anaemia and β -thalassaemias⁵⁶. Although the common SNP in the *BCL11A* gene has only a modest effect, strong perturbation of the gene in cells results in half of the cell's haemoglobin being HbF.

Type 2 diabetes. Studies have identified 39 loci so far in this disorder, which affects 300 million people worldwide⁵⁷. Notably, genes previously implicated in glucose regulation based on biochemical studies do not seem to be associated with type 2 diabetes, but are associated with fasting glucose levels. The pathophysiology thus probably involves different molecular mechanisms. Many of the genes also point to insulin secretion rather than insulin resistance as the primary cause.

Autoimmune diseases. Studies have found ~100 loci related to autoimmune diseases such as type 1 diabetes, rheumatoid arthritis and coeliac disease. These studies point to many loci that have roles across multiple autoimmune diseases and probably involve fundamental regulatory pathways, as well as many that are disease specific.

Height. Adult height is a classic polygenic trait with high heritability. A recent analysis of 180,000 individuals identified 180 loci⁵⁸, with many showing multiple distinct alleles. With the large number of loci and new analytical tools, various biological pathways have been implicated, such as TGF- β signalling.

Kidney disease. GWAS for two common renal disorders revealed variants in *APOL1* that are common in African chromosomes but absent in European chromosomes, which account for much of the increased risk of kidney disease in African-Americans. The variants seem to have reached high frequency in Africa because they protect carriers from the deadliest subtype of African sleeping sickness.

Psychiatric disorders. Studies of psychiatric diseases have been relatively limited in scope, and much less is known than for other diseases. Genotyping studies have identified both common variants (in bipolar disorder and schizophrenia^{59,60}) and rare deletions (in autism^{46,61} and schizophrenia⁶²). Each class of variants so far accounts for a few per cent of the genetic variance, but analyses indicate that both classes will have a major role in a highly polygenic aetiology^{45,47,60,63}. Given our near-total ignorance of the underlying cellular pathways, larger genetic studies are essential.

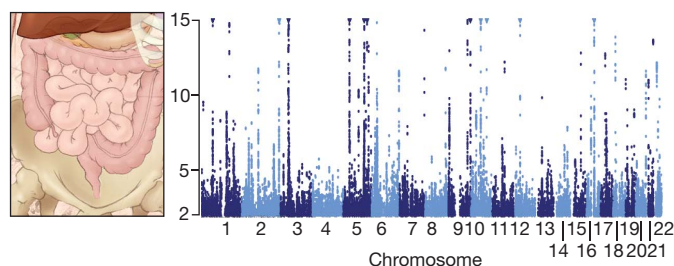


Figure 3 | Disease association maps. Geneticists can now test the association between a common disease and millions of individual genetic variants. The figure shows a 'Manhattan plot' from a study of Crohn's disease, a form of inflammatory bowel disease. For each variant across the genome, the height reflects its correlation with disease (measure by $\log_{10}(\text{significance})$). The Manhattan plot reveals 71 'skyscrapers', corresponding to regions associated with Crohn's disease. Image courtesy of B. Wong (ClearScience).

Pharmacogenetics. Association studies have revealed genetic factors underlying hypersensitivity to the antiretroviral drug abacavir, drug-induced myopathies associated with cholesterol-lowering drugs, cardiovascular risks in patients receiving the antiplatelet drug clopidogrel and variations in metabolic clearance of the anticoagulant warfarin.

eQTLs. GWAS have also proved powerful for studying gene regulation, by using cellular gene expression as a phenotype in its own right. Following early studies associating variation in transcript levels with nearby common variants^{64,65}, studies have identified thousands of quantitative trait loci affecting gene expression (eQTLs), both in *cis* and *trans*.

Notwithstanding the moderate effect sizes in disease studies, investigators have been able to definitely implicate specific mutations in humans, by combining eQTL studies and transgenic animal models⁶⁶. Although many affect coding sequences, a larger proportion seems to lie in non-coding regions and may affect gene regulation, consistent with the abundance of functional non-coding sequences in the genome. A stunning example is the region around the gene encoding the cell-cycle regulator CDKN2A/B, in which distinct non-coding regions are associated with a dozen different diseases, including diabetes, heart disease and various cancers. Some of the causal regulatory variants have been discovered, but most remain mysterious. These findings call for improved methods to understand the function of non-coding regions and seem likely to reveal new mechanisms of gene regulation.

Genetic architecture of disease

Despite this success, the results have provoked some handwringing in the scientific community and beyond because initial studies often explained only a small proportion of the heritability (defined as the additive genetic variance). The so-called ‘missing heritability’ has provoked renewed interest in the ‘genetic architecture’ of human disease and traits—a topic that was the subject of much debate early in the twentieth century. The explanation will surely involve multiple contributions.

First, it is becoming clear that the heritability due to common variants is greater than initially appreciated. With larger GWAS, the heritability explained has continued to grow, reaching 20–25% for various diseases and traits (Table 1). Moreover, current estimates substantially understate the actual role of common variants for two reasons. One reason is that current GWAS miss many common variants of lower frequency (1–10%), because existing genotyping arrays often lack useful proxies. Many disease-related alleles probably fall into this frequency class, which is enriched for variants under mildly deleterious selection. New genotyping arrays based on the 1000 Genomes Project should be able to capture these common variants. Another reason is that GWAS also miss many common variants of smaller effect, due to limited sample size and stringent statistical thresholds imposed to ensure reproducibility. Recent efforts have sought to infer the contribution of loci that fall just short of statistical significance⁶⁷. Beyond, there are surely many more common variants with still smaller effects (the standing variation expected under Darwinian evolution): although their individual contributions may be too small to ever detect with feasible sample sizes, they may collectively explain a significant fraction of heritability. Elegant indirect analyses indicate that common variants must account for >55% of heritability for height and >33% for schizophrenia^{60,68}.

Table 1 | GWAS for common diseases and traits

Phenotype	Number of GWAS loci	Proportion of heritability explained (%) [*]
Type 1 diabetes	41	~60
Fetal haemoglobin levels	3	~50
Macular degeneration	3	~50
Type 2 diabetes	39	20–25
Crohn’s disease	71	20–25
LDL and HDL levels	95	20–25
Height	180	~12

HDL, high-density lipoprotein; LDL, low-density lipoprotein.

^{*} Fraction of heritability explained is calculated by dividing the phenotypic variance explained by variants at loci identified by GWAS by the total heritability as inferred from epidemiological parameters. (See Supplementary Bibliography.)

Second, rare variants of larger effect will also play an important part in common diseases, although their role has barely been explored. Studying them requires sequencing protein-coding (or other) regions in genes to identify those in which the aggregate frequency of rare variants is higher in cases than controls. So far, studies have focused on candidate genes implicated through Mendelian diseases, mouse mutants, biochemical pathways or GWAS studies. Early studies reported findings at *MC4R* in extreme, early-onset obesity and *PPARG* in insulin resistance (both of which also have common variants affecting the traits). Recent important examples include rare variants in six candidate genes affecting lipid levels⁶⁹ and three candidate genes affecting blood pressure⁷⁰. A study of a GWAS region associated with HbF levels implicated the *MYB* gene, explaining additional heritability beyond the common variants⁷¹.

Whether rare variants will reveal many new genes must await systematic whole-exome sequencing. Given the background rate of rare variants (~1% per gene), many thousands of samples will be needed to achieve statistical significance. Similarly, the total heritability due to rare variants is still unclear. Although the inferred effect sizes are larger, the overall contribution to the heritability may still be small due to their low frequency. For example, only ~1/400 of the heritability is explained by rare variants at each of the three loci affecting blood pressure. Whatever their contribution to heritability, rare variants of large effect will be valuable by enabling direct physiological studies of pathways in human carriers.

Finally, some of the missing heritability may simply be an illusion. Heritability is estimated by applying formulae for inferring additive genetic effects from epidemiological data. The estimates may be inflated because the methods are not very effective at excluding the (nonlinear) contributions of genetic interactions or gene-by-environment interactions, which are likely to be significant.

Biological mechanisms versus risk prediction

It is important to distinguish between two distinct goals. The primary goal of human genetics is to transform the treatment of common disease through an understanding of the underlying molecular pathways. Knowledge of these pathways can lead to therapies with broad utility, often applicable to patients regardless of their genotype. The past decade has seen remarkable progress towards identifying disease genes and pathways, with greater advances ahead.

Some seek a secondary goal: to provide patients with personalized risk prediction. Although partial risk prediction will be feasible and medically useful in some cases, there are likely to be fundamental limits on precise prediction due to the complex architecture of common traits, including common variants of tiny effect, rare variants that cannot be fully enumerated and complex epistatic interactions, as well as many non-genetic factors.

The road ahead

The discovery of more than 1,100 loci within a few years is an excellent start, but just a start. Over the next decade, we need genetic studies of tens of thousands of patients for most common diseases, with appropriate combinations of GWAS and sequencing. In turn, intensive functional studies will be required to characterize the genes and pathways, and to construct animal models that mimic human physiology. Importantly, complete explanation of a disease is not required for progress.

Medicine: cancer

The view from 2000

With the establishment by the early 1980s that cancer results from somatic mutations, Dulbecco declared in an influential article in 1986 that sequencing the human genome was a critical priority, saying “We have two options: either to try to discover the genes important in malignancy by a piecemeal approach, or to sequence the whole genome of a selected animal species”⁷². By 2000, ~80 cancer genes involved in solid tumours had been discovered, most through viral oncogenes and transformation assays and the remainder through positional cloning of inherited cancer

syndromes and somatically deleted regions in cancer. (Many additional genes were found in blood cancers, where translocations could be readily visualized and cloned.)

A decade later, cancer gene discovery is being driven by systematic genome-wide efforts, involving a powerful combination of DNA sequencing, copy-number analysis (using genotyping arrays developed for GWAS), gene expression analysis and RNA interference. The number of cancer genes in solid tumours has nearly tripled to ~230, revealing new biological mechanisms and important therapeutic leads (see ref. 73 and <http://www.sanger.ac.uk/genetics/CGP/cosmic>). More generally, systematic efforts are beginning to parse the vast heterogeneity of cancer into more homogeneous groupings based on mechanism.

Early sequencing efforts

Given the limited capacity of capillary-based sequencing, initial sequencing efforts focused on targeted gene sets, such as kinases in signalling pathways underlying cell growth; they soon hit pay dirt. (1) *BRAF* mutations were discovered in >50% of melanomas⁷⁴. Pharmaceutical companies raced to develop inhibitors of RAF and MEK, a gene downstream of RAF. Initial results were disappointing, but recent clinical trials (using more potent inhibitors and studying tumours carrying relevant mutations) have shown response rates exceeding 80%. (2) *PIK3CA* mutations were discovered in >25% of colorectal cancers⁷⁵; pharmaceutical programmes are at an earlier stage. (3) *EGFR* mutations were discovered in 10–15% of lung cancers and predicted which patients would respond to gefitinib and erlotinib, drugs that had had only patchy efficacy⁷⁶. Such treatment soon became the standard-of-care for patients with the relevant mutation and has been shown to extend life.

An early exome-wide sequencing study of glioblastoma found a new class of cancer gene involved in basic cellular metabolism: a recurrent mutation in *IDH1* alters the active site of the enzyme isocitrate dehydrogenase⁷⁷, causing it to aberrantly generate an 'oncometabolite', 2-hydroxyglutarate. Pharmaceutical companies are already working towards the development of inhibitors of the neo-enzyme.

Microarray-based studies

Genotyping arrays allowed genome-wide, high-resolution analysis of amplifications and deletions. A recent genomic study of >3,000 tumours across 26 cancer types catalogued more than 150 recurrent focal copy-number alterations⁷⁸; only one-quarter contain known cancer genes, indicating that many more remain to be discovered. Studies of specific events have identified many new cancer genes. Amplifications revealed an entirely new class consisting of transcription factors necessary for lineage-specific survival (*MITF* in melanoma, *NKX2.1* in lung cancer and *SOX2* in oesophageal cancer)⁷⁹. Recurrent deletions in paediatric acute lymphoblastic leukaemias were found in *PAX5*, *IKZF1* and other regulators of lymphocyte differentiation.

Gene-expression arrays similarly revealed new mutational targets, including functional translocations involving one of several ETS transcription factors in >50% of prostate tumours⁸⁰, disproving the long-held belief that translocations have a major role in blood cancers, but not epithelial solid tumours. Translocations involving *ALK* have since been found in some lung cancers, and new pharmaceuticals directed against *ALK* are already showing impressive results in clinical trials⁸¹.

Genome-wide expression analysis has also had a central role in classifying cancers based on their molecular properties^{82,83}, rather than anatomic sites. Studies have revealed distinctive subtypes and shed light on metastatic potential⁸⁴. Expression signatures are already used in the clinic to predict which breast cancer patients will benefit most from adjuvant chemotherapy after surgery. More generally, gene expression analysis has become a routine aspect of both basic and clinically oriented discovery research. Expression signatures have also been proposed as a lingua franca for connecting the molecular mechanisms of drugs, genes and diseases⁸⁵, sometimes revealing new ways to use old drugs⁸⁵.

Genomic studies have been integrated with genome-wide RNA-interference-based screens to identify genes that are both genomically

amplified and essential for cell viability in specific cancer cell lines. Examples include *IKBKE* in breast cancer, *CDK8* in colorectal cancer and the nuclear export protein *XPO4* in hepatocellular cancer⁸⁶. Genomically characterized cell lines can also be screened to identify 'synthetic lethals'—that is, genes essential only in the presence of particular recurrent cancer mutations—such as *PLK1*, *STK33* and *TBK1*, required in the setting of *KRAS* mutations.

Genome-wide sequencing

With massively parallel sequencing, attention has now focused on comprehensive exome-wide and genome-wide sequencing in large numbers of samples (Fig. 4). In acute myelogenous leukaemia and clear-cell ovarian cancer, recurrent mutations in *DNMT3A* and *ARID1A*, respectively, point to epigenomic dysregulation^{87,88}. Studies of tumours from ~40 multiple myeloma patients have uncovered frequent mutations in genes not previously known to have a role in cancer (such as *DIS3* and *FAM46C*), implicating novel pathways such as protein translation and homeostasis, as well as NF- κ B activation⁸⁹. Whole-genome sequencing of prostate cancers has shed light on the origins of tumour rearrangements.

The road ahead

The ultimate goal is to markedly decrease death from cancer. To guide therapeutics, we must develop over the next decade a comprehensive catalogue of all genes that are significant targets of somatic alteration in all human cancer types, all animal models and all cancer cell lines.

The patterns of mutated genes will: (1) define direct drug targets in some cases; (2) identify cellular pathways and synthetic lethal interactors to target in others; (3) direct the creation of animal models; (4) allow chemical screening against cancer cells with defined molecular mechanisms; and (5) guide the design of human clinical trials. We must also chart all ways in which tumours develop resistance to specific interventions in patients. Ideally, we should obtain this information from pre-clinical studies, so that we can plan countermeasures, even as we develop a drug. Effective cancer treatment will surely require combination therapies, like those used against HIV, that target multiple vulnerabilities to markedly diminish the chance of resistance.

Projects such as The Cancer Genome Atlas and the International Cancer Genome Consortium plan to study ~500 tumours per cancer type, but these goals will need to be dramatically expanded. As genomics permeates clinical practice, we should create a mechanism by which all cancer patients can choose to contribute their genomic and clinical data to an open collaborative project to accelerate biomedical progress.

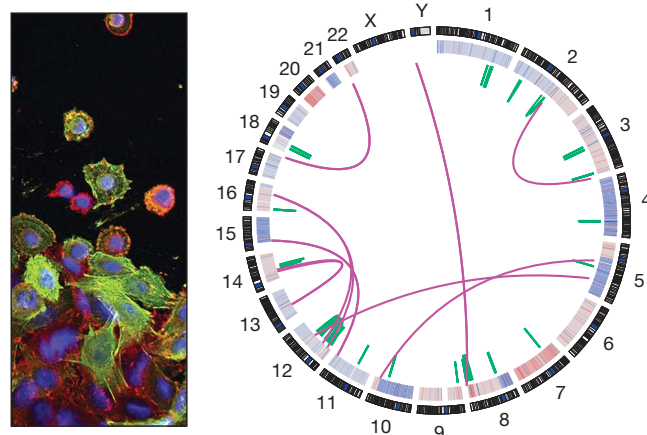


Figure 4 | Cancer genome maps. Whole-genome sequencing has provided powerful new views of cancers. The left panel shows an image of colon cancer (Wellcome Trust). The right panel shows the genome of a colon cancer sample (Broad Institute), including interchromosomal translocations (purple), intrachromosomal translocations (green) and amplifications and deletions (red and blue, on the inner ring). Individual nucleotide mutations are not shown.

Human history

The view from 2000

Well before 2000, studies of genetic variation at handfuls of loci such as mitochondrial DNA and blood groups across worldwide populations had given rise to an intellectual synthesis according to which modern humans arose in Africa, dispersed from the continent in a single migration event 50,000–100,000 years ago, replaced resident archaic human forms, and gave rise to modern populations largely through successive population splits without major mixture events⁹⁰. It was difficult, however, to reconstruct the details of these events from the limited data. In addition, little was known about the role that positive selection may have played in shaping the biology of human populations as they migrated and expanded.

A decade later, genomic data have radically reshaped our understanding of the peopling of the globe, yielding a vastly richer picture of population mixing and natural selection. These studies have been made possible by the growth in catalogues and maps of genetic variation among human populations, as well as differences with our closest relatives, such as Neanderthal and chimpanzee.

Population mixture in human history

Shifting the focus from stories based on individual loci (such as ‘mitochondrial Eve’) to large collections of genetic markers has provided powerful new insights. It is now clear that the migration of humans out of Africa was more complicated than previously thought, and that human history involved not just successive population splits, but also frequent mixing.

There is now strong genetic evidence, for example, showing that south Asians are the product of an ancient mixture. European populations have mixed so extensively with their neighbours that their genes mirror geography, rather than reflecting the paths of human migrations or language families.

Most strikingly, genome analysis showed that anatomically modern humans mixed with the Neanderthals. Europeans and Asians (but not Africans) have all inherited 1–4% of their genome from Neanderthals, indicating gene flow in the Middle East on the way out of Africa⁹.

Positive selection in the last 10,000 years

By studying dense collections of genetic markers, it has become possible to spot the signs left of recent positive natural selection—even without knowing the specific trait under selection. One such signature is a high-frequency, long-range haplotype, which results when an advantageous genetic variant rapidly sweeps through a population and carries along neighbouring variants⁹¹. Analysis of data from the HapMap has revealed at least 300 genomic regions that have been under positive selection during the past ~5,000–30,000 years. New methods have narrowed these signals to small regions, often with only a single gene, and sometimes implicated specific candidate variants⁹² (Fig. 5).

Although much work will be needed to decipher each gene’s unique story, the implicated genes are already beginning to point to specific processes and pathways. Many encode proteins related to skin pigmentation, infectious agents, metabolism and sensory perception⁹². In Europe, powerful selection around the dawn of agriculture favoured a regulatory variant that causes lifelong expression of lactase, the enzyme required to digest milk; a similar mutation was independently selected in cattle-herding groups in East Africa. In West Africa, strong selection for a gene encoding a receptor for the Lassa fever virus may indicate a resistance allele. Comparisons of nearby populations living under different conditions can be particularly informative. Tibetans, a population living at 14,000 feet, and Han Chinese are closely related, but show striking population differentiation at a locus encoding a protein involved in sensing oxygen levels⁹³.

Positive selection in human speciation

A favourite question of philosophers and scientists alike is ‘what makes us uniquely human?’ The human and chimpanzee genome sequences

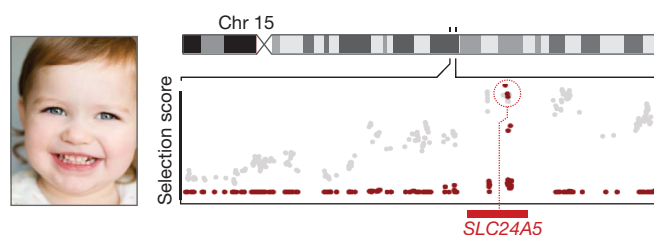


Figure 5 | Positive selection maps. Genetic variation patterns across the genome can help to reveal regions that have undergone strong positive selection during human history. The figure shows a region of 1 million bases on chromosome 15 that has been under strong selection in European populations. The initial analysis (grey) provides diffuse localization, whereas fine-structure mapping (red) pinpoints the signal to a gene (*SLC24A5*) known to affect skin colour. Image courtesy of iStock Photo.

made it possible to give a definitive, if unsatisfying answer: we can now enumerate the ~40 million genetic differences⁷. Unfortunately, the vast majority of them represent random drift.

The best clues as to which changes drove human evolution have come from methods to detect ancient positive selection. At several hundred loci termed ‘human accelerated regions’ (HARs), the rate of nucleotide change since the divergence from the common ancestor of human and chimpanzee is exceptionally high relative to the rate of change over the previous 100 million years⁹⁴. HAR1 is part of a non-coding RNA that is expressed in a region of the brain that has undergone marked expansion in humans. HAR2 includes a transcriptional enhancer that may have contributed towards the evolution of the opposable thumb in humans. Similarly, the *FOXP2* gene has undergone accelerated amino-acid substitution along the human lineage. Because null alleles of *FOXP2* affect language processing, it has been suggested that these changes may be related to our acquisition of language. Of course, large signals of accelerated evolution represent only a piece of the puzzle: many critical changes surely involved only isolated nucleotide changes. Identifying these changes will require combining insights from both genotypic and phenotypic differences. Several dozen candidates have been suggested.

The road ahead

The ultimate goal is to use genomic information to reconstruct as much as possible about the salient events of human history. This includes a complete accounting of the structure of the ancestral human population in Africa; the subsequent population dispersals and their relationship to landmarks such as the spread of agriculture; gene flow with archaic hominins both before and after the Out-of-Africa migration; and the impact of positive natural selection in recent and ancient times.

Over the next decade, we should assemble large-scale genomic databases from all modern human populations, hominin fossils and great ape relatives. New laboratory techniques will be needed to infer and test the functional role of human variations. Advances in statistical methodology will be needed, including better ways to date events and exploit haplotype information to infer common ancestry.

New frontiers

Twenty-five years ago, biologists debated the value of sequencing the human genome. Today, young scientists struggle to imagine the nature of research in the antediluvian era, before the flood of genomic data. Genomics has changed the practice of biology in fundamental ways. It has revealed the power of comprehensive views and hypothesis-free exploration to yield biological insights and medical discoveries; the value of scientific communities setting bold goals and applying teamwork to accomplish them; the essential role of mathematics and computation in biomedical research; the importance of scale, process and efficiency; the synergy between large-scale capabilities and individual creativity; and the enormous benefits of rapid and free data sharing.

Yet, this is only a step towards transforming human health. We must now extend these principles to new frontiers. Our goal should be to

dramatically accelerate biomedical progress by systematically removing barriers to translational research and unleashing the creativity of a new generation.

Within genomics, we must complete biomedicine's 'periodic table'⁵¹. This will take at least another decade, through systematic efforts to define the components such as those described above. Connecting these components fully with disease will require something like an international 'One Million Genomes Project', analysing well-annotated patient samples from many disorders.

We must also apply systematic approaches more broadly. Some specific goals are given in the following paragraphs.

Modular cell biology

Just as it was once inconceivable to possess complete catalogues of cellular components, it seems quixotic today to seek a comprehensive picture of cellular circuitry. Yet, it is time to turn systematic attention to this next level of organization. Cellular circuitry is not infinitely complex. It is organized around a limited repertoire of modules⁹⁵, whose reuse is fundamental to evolvability. These modules include protein complexes, *cis*-regulatory circuits, metabolic pathways, and signalling networks, each involving tightly coupled cores with hierarchies of condition-dependent interactions⁹⁶. In yeast, we can already begin to glimpse the basic modular organization that controls environmental responses. In mammals, the picture will be far more complex but the number of fundamental modules is likely to be tractable—perhaps a few thousand. The goal of a complete catalogue of cell modules will involve many challenges. Conceptually, we will need diverse ways to infer and test candidate modules (for example, correlations in gene expression, protein modification and evolutionary retention and by systematic perturbation). Technically, we will need powerful platforms—'cell observatories', so to speak—including systematic reagents for modulating components and interactions; new instruments for single-cell measurements; analytical methods to derive mechanistic models; and access to many cell types⁹⁶.

Cell programming

We must learn to program cells with the same facility with which we program computers. The past decade has set the stage. Yamanaka's stunning discovery that adult cells can be re-programmed into pluripotent cells has inspired screens for particular gene cocktails to induce other transformations. Independently, a cadre of creative young synthetic biologists have begun to invent new cellular circuits. The key challenges ahead include developing a general recipe to trans-differentiate any cell type into any other, and general combinatorial tools that make it is easy to create circuits activated only in a specific cellular state. Cell programming will draw inspiration from native cellular modules and in turn provide tools to study them.

Chemical biology and therapeutic science

Accelerating treatments for disease will require a renaissance in therapeutic science. The pharmaceutical industry will remain the locus of drug development, but it rightly focuses on proven methods and commercial markets. Academia must become a hotbed for heterodox approaches that combine creativity and scale, exploit genomic approaches and targets, and empower a new generation of scientists. Key goals include developing large libraries of small molecules whose chemical properties favour selectivity, potency and rapid optimization⁹⁷; powerful phenotypic assays using genomic signatures and other general approaches that make it possible to find modulators for any cellular process^{98,99}; and systematic methods that can rapidly and reliably find the protein targets and mechanisms of action of 'hits'. Ultimately, we should aim for a comprehensive arsenal of small-molecule modulators for all cellular targets and processes to probe physiology and test therapeutic hypotheses.

Medical revolutions require many decades to achieve their full promise. Genomics has only just begun to permeate biomedical research: advances must proceed through fundamental tools, basic discoveries, medical studies, candidate interventions, clinical trials, regulatory approval and widespread

adoption. We must be scrupulous not to promise the public a pharmacopoeia of quick pay-offs. At the same time, we should remain unabashed about the ultimate impact of genomic medicine, which will be to transform the health of our children and our children's children.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). **The draft sequences reported in refs 1 and 2 provided the first comprehensive look at ~90% of the human genome; the finished sequence in ref. 3 increased the completeness to >99% and the accuracy to >99.999%, providing a solid foundation for biomedicine.**
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002). **Comparison with the mouse genome led to the discovery that the vast majority of functional sequence in the human genome does not encode protein.**
5. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
6. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
7. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
8. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
9. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
10. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324**, 1190–1192 (2009).
11. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
12. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
13. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
14. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
15. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
16. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**, 907–909 (2007).
17. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).
18. Guttman, M. *et al.* *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.* **28**, 503–510 (2010).
19. Yassour, M. *et al.* *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 3264–3269 (2009).
20. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*. doi:10.1073/pnas.1017351108 (27 December, 2010).
21. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* **104**, 19428–19433 (2007).
22. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
23. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
24. Pennacchio, L. A. *et al.* *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
25. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
26. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
27. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
28. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
29. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
30. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
31. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
32. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
33. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
34. Zappulla, D. C. & Cech, T. R. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 217–224 (2006).

35. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
36. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
37. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
38. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
39. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
40. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
41. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
42. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
- References 40–42 laid the foundation for genetic studies of common disease, which have so far identified more than 1,100 loci associated with diseases.**
43. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
44. McCarroll, S. A. Copy number variation and human genome maps. *Nature Genet.* **42**, 365–366 (2010).
45. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
46. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
47. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
- References 45–47 revealed an important role of rare genetic deletions in psychiatric diseases.**
48. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
49. Bilgüvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207–210 (2010).
50. Habashi, J. P. *et al.* Losartan, an AT1 antagonist, prevents aortic aneurysm in a mouse model of Marfan syndrome. *Science* **312**, 117–121 (2006).
51. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
52. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
53. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
54. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
55. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–1125 (2010).
56. Uda, M. *et al.* Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β -thalassaemia. *Proc. Natl Acad. Sci. USA* **105**, 1620–1625 (2008).
57. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genet.* **42**, 579–589 (2010).
58. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
59. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747 (2009).
60. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
61. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
62. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25**, 528–535 (2009).
63. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
64. Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. *Nature Genet.* **32**, 432–437 (2002).
65. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
66. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
67. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42**, 570–575 (2010).
68. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
69. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nature Genet.* **37**, 161–165 (2005).
70. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genet.* **40**, 592–599 (2008).
71. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genet.* **42**, 1049–1051 (2010).
72. Dulbecco, R. A turning point in cancer research: sequencing the human genome. *Science* **231**, 1055–1056 (1986).
73. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
74. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
- The discovery of *BRAF* mutations in melanoma has led to new drugs for melanoma with high response rates.**
75. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
76. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
77. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
78. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
79. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
80. Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
81. Kwak, E. L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* **363**, 1693–1703 (2010).
82. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
83. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
84. van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
85. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
86. Boehm, J. S. *et al.* Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell* **129**, 1065–1079 (2007).
87. Ley, T. J. *et al.* *DNMT3A* mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).
88. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
89. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* doi:10.1038/nature09837 (in the press).
90. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and geography of Human Genes* 518, 541 (Princeton Univ. Press, 1994).
91. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
92. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
93. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
94. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006).
95. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
96. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic, 2006).
97. Clemons, P. A. *et al.* Small molecules of different synthetic and natural origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl Acad. Sci. USA* **107**, 18787–18792 (2010).
98. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
99. Yarrow, J. C., Feng, Y., Perlman, Z. E., Kirchhausen, T. & Mitchison, T. J. Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Throughput Screen.* **6**, 279–286 (2003).
100. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This review reflects collective ideas, insightful conversations and contributions shared by many colleagues at the Broad Institute and elsewhere. In particular, I wish to express my gratitude to D. Altshuler, J. Baldwin, B. Bernstein, B. Birren, C. Burge, F. Collins, M. Daly, M. DePristo, E. Eichler, A. Futreal, L. Garraway, T. Golub, E. Green, C. Gunter, M. Guyer, M. Guttman, D. Haussler, E. Hechter, J. Hirschhorn, D. Hung, D. Jaffe, S. Kathiresan, L. Kruglyak, E. Lieberman, R. Lifton, K. Lindblad-Toh, S. McCarroll, A. Meissner, T. Mikkelsen, R. Myers, R. Nicol, C. Nusbaum, L. Pennacchio, R. Plenge, A. Regev, D. Reich, J. Rinn, P. Sabeti, V. Sankaran, S. Schreiber, P. Sklar, M. Stratton, H. Varmus, P. Visscher, A. Wolf and O. Zuk. I also thank B. Wong and L. Gaffney for assistance with figures.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to E.S.L. (lander@broadinstitute.org).

A decade's perspective on DNA sequencing technology

Elaine R. Mardis¹

The decade since the Human Genome Project ended has witnessed a remarkable sequencing technology explosion that has permitted a multitude of questions about the genome to be asked and answered, at unprecedented speed and resolution. Here I present examples of how the resulting information has both enhanced our knowledge and expanded the impact of the genome on biomedical research. New sequencing technologies have also introduced exciting new areas of biological endeavour. The continuing upward trajectory of sequencing technology development is enabling clinical applications that are aimed at improving medical diagnosis and treatment.

The sequencing of the Human Reference Genome, announced ten years ago, provided a roadmap that is the foundation for modern biomedical research. This monumental accomplishment was enabled by developments in DNA sequencing technology that allowed data production to far exceed the original description of Sanger sequencing¹. Moving forward in the genomic era in which we now find ourselves, new (or 'next generation') DNA sequencing technology is enabling revolutionary advances in our understanding of health and disease. In essence, sequencing technology is the engine that powers the car that allows us to navigate the human genome roadmap. As that engine becomes ever more powerful, so will the questions we can ask and answer about the geography of our genetic landscape.

Of course, a car with only an engine is unworkable; as such, DNA sequencing technology provides an integral part of a larger system, one with multiple components that must be properly matched in order to achieve high throughput and efficiency. It has essentially never been as 'easy' as simply buying sequencing instruments, plugging them in, and generating data. We need the raw materials, such as fuel (DNA), sparks to ignite the fuel (reagents), mechanical parts to translate fuel and ignition into movement (robotics) and direction (bioinformatics), all working in a carefully engineered balance, and a driver (genome centre) to steer the automobile quickly and efficiently to the desired destination (biological understanding). By inference, as this 'engine' has achieved ever increasing horsepower, the supporting components have evolved to match its output with corresponding levels of performance, and new or completely revised components have been added as required.

In 2001, the technology that sequenced the human genome was based on capillary electrophoresis of individual fluorescently labelled Sanger sequencing reaction products. Each instrument could detect 500–600 bases from each of 96 reactions in around ten hours, with 24-hour unattended operation producing 115 kbp (thousand base pairs) per day. Because of the increased scale required for the Human Genome Project, genome centres had developed a robust, highly automated and inexpensive preparatory process to feed their capillary sequencers. Once the data were produced, mature analysis software was applied to analyse the sequencing reads (each a ~500-bp sequence of A, C, G, T), then to assemble reads that shared sequence identity, reproducing that region of the genome. After assembly, each genomic region was further analysed to identify genes, repeat elements and other features. As the 'drivers' of these sequencing pipelines, genome centres could dial up capacity by increasing the amount of hardware used in the preparatory and sequencing

processes, because sequence production, not sequence analysis, was rate limiting.

As I will describe, the ensuing ten years has been marked by dramatic improvements in sequencing technology that have catapulted sequencing to the forefront of biological experimentation and have revolutionized the way that we approach genome-wide questions. One consequence of this revolution has been the coincident revitalization of bioinformatics, predominantly in development efforts aimed at data analysis and interpretation. Taken together, these unprecedented sequencing and analysis capabilities have inspired new areas of enquiry, have solved major questions about the regulation, variability and diaspora of the human genome, and have introduced a genomic era in medical enquiry and (ultimately) practice that will bring about the originally envisioned impact of the Human Genome Project.

Massively parallel sequencing

The first five years following the Human Genome Project provided further definition and annotation of the human genome sequence by comparative genomics; the sequencing of several model organism genomes—such as mouse², rat³, chicken⁴, dog⁵, chimpanzee⁶, rhesus macaque⁷, duckbill platypus⁸ and cow⁹—provided information about highly conserved genomic elements that are likely to be functional owing to their conservation. These genomes were largely produced by conventional methods, including Sanger-based capillary sequencing. Starting in 2005, a variety of new 'engines' for DNA sequencing that were radically different from the capillary sequencers used to sequence the human and model organism genomes became available from several different manufacturers (Fig. 1). These new engines were 'turbo-charged' by several orders of magnitude compared to their predecessors, because the basic mechanisms for data generation had changed radically, producing far more sequence reads per instrument run and at a significantly lower expense. The availability of multiple commercially available instruments alone represented a paradigm shift from the previous decade, where a single capillary instrument produced by Applied Biosystems dominated the market. Many of these innovative approaches were initially developed with National Institutes of Health (NIH) funding through the 'Technology development for the \$1,000 genome' program (<http://www.genome.gov/11008124#al-4>) introduced during Francis Collins' directorship at the National Human Genome Research Institute (NHGRI).

Since the introduction of these platforms, the past five years have been marked by fierce competition between their manufacturers to greatly

¹The Genome Center at Washington University School of Medicine, Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63108, USA.

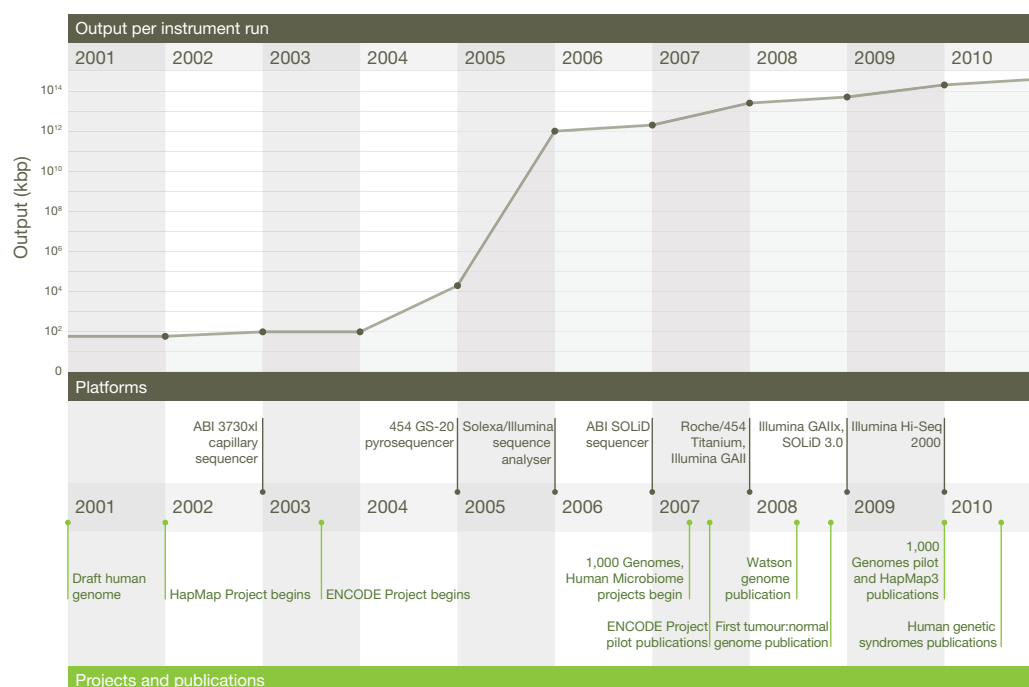


Figure 1 | Changes in instrument capacity over the past decade, and the timing of major sequencing projects. Top, increasing scale of data output per run plotted on a logarithmic scale. Middle, timeline representing major

milestones in massively parallel sequencing platform introduction and instrument revisions. Bottom, the timing of several projects and milestones described in the text.

increase the amount of sequence output per run, to increase read lengths (the number of nucleotides per sequence read), to lower costs, and to improve base-calling accuracy for their instruments. Much like a buyer of a new car, genome centres have taken each new 'massively parallel' (see below) instrument for a 'test drive' to explore their performance and to understand their strengths and weaknesses in terms of data quality, associated production processes, and actual cost to operate. (This last includes personnel, consumables, additional instrumentation to operate, amortized equipment rate, and informatics infrastructure.) Their collective efforts to vet each new technology and report their findings by scientific presentations, press releases, word of mouth, and peer-reviewed manuscripts have effectively fuelled the rivalry and competition among the instrument vendors and have resulted, not surprisingly, in both winners and losers in this commercial sector.

This so-called 'massively parallel' sequencing technology differs significantly from Sanger capillary sequencing. Although each instrument is distinctly different in its specifics, as detailed in several reviews^{10,11} (see also Table 1), all massively parallel devices share certain attributes, as

follows. First, the initial preparatory steps are fewer and simpler to perform than for Sanger sequencing. Instead of a bacterial cloning step followed by DNA isolation, massively parallel sequencing begins with the production of a library formed by ligating platform-specific synthetic DNAs (adapters) onto the ends of the fragment population to be sequenced. Second, all platforms require the library fragments to be amplified on a solid surface (either a glass slide or a microbead) by a polymerase-mediated reaction that produces many copies of each single library fragment. Amplification is needed so that the ensuing sequencing reactions produce sufficient signal for detection by the instrument's optical system. However, this step also provides a source of sequencing error that is perpetuated through the downstream processes, because polymerases are never 100% accurate. Third, these instruments perform sequencing reactions as an orchestrated series of repeating steps that are performed and detected automatically. The specifics of the DNA sequencing reaction are different for each platform, emphasizing the amazing range of innovation in chemistry, molecular biology and engineering required to produce sequence information from hundreds

Table 1 | Sequencing platform comparison

	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000	Pacific Biosciences RS
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface	NA (single molecule detection)
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase-mediated incorporation of end-blocked fluorescent nucleotides	Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3' end of oligonucleotide	Chemical cleavage of fluorescent dye and 3' blocking group	NA (fluorescent dyes are removed as part of PPi release on nucleotide incorporation)
Error model	Substitution errors rare, insertion/deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/deletion errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp	>1,000 bp

Comparison of commercially available next generation platforms (Roche/454, Life Technologies and Illumina) and a single molecule platform (Pacific Biosciences), illustrating the similarities and differences in these technologies, according to several metrics. NA, not applicable; PPi, pyrophosphate.

*emPCR (emulsion PCR) is a bulk amplification process whereby library fragments are combined with beads and PCR reactants in an oil emulsion that allows *en masse* amplification of millions of bead-DNA combinations in a single tube.

of thousands to hundreds of millions of DNA molecules simultaneously. For example, the Roche/454 instrument detects each polymerase-catalysed nucleotide incorporation event by a downstream series of reactions that produce light ('pyrosequencing'), initiated by the pyrophosphate molecules released on nucleotide incorporation. The Life Technologies SOLiD uses a unique DNA ligase-mediated process that, through multiple rounds of template-directed ligation, sequences each nucleotide twice. The Illumina sequencer incorporates fluorescently labelled nucleotides that are chemically blocked such that only one nucleotide incorporation event occurs per fragment population per sequencing cycle. Regardless of the details, massively parallel sequencing reactions are distinguished by the fact that they occur in a nucleotide-by-nucleotide stepwise fashion, rather than by discrete separation and detection (in a 96-at-a-time fashion) of already produced Sanger sequencing reaction products on a capillary instrument. The fourth shared feature of these systems is the ability to obtain sequence information from both the ends of the DNA fragments comprising the sequencing library. Depending on the instrument system and the library construction approach used, one can either sequence at both ends of linear fragments ('paired end sequencing') or from both ends of previously circularized fragments ('mate pair sequencing').

Paired end sequencing libraries are the standard means by which human genomes are sequenced, because they are straightforward to make and require a small amount of DNA. Mate pair libraries, by contrast, are quite DNA expensive owing to the low yield of circularization of large DNA molecules (a yield that diminishes proportionally with increasing length of the DNA molecules used). However, mate pair libraries provide valuable information about larger structural events because they sample DNA sequence over a larger distance (1.5–20 kilobases, kb) than do paired end libraries (~300–500 bp). The benefit of obtaining sequence data from both ends of library fragments in human genome sequencing is obvious when one considers the highly repetitive nature of the genome. Explicitly, aligning at least one end read of a pair uniquely onto the reference sequence provides sufficient certainty that the read pair is uniquely mapped to its locus of origin. Conversely, aligning short, single end or 'fragment' reads to the genome results in a higher proportion of non-unique placements—reads that cannot be used for variant discovery. As described later, the other value of paired end data lies in its use for discovering structural variation in the genome.

Although massively parallel platforms have significantly affected our ability to study the human genome and to better understand its variability in a multitude of contexts, these technologies have required profound changes to the data analysis pipelines that previously had been so straightforward. In particular, the new sequencing engines have introduced data analysis challenges owing to the massive scale of the data to be analysed, the significant decrease in the read length, and in the dramatically different error profiles of each read type, when compared to those of capillary data. These new challenges have resulted in a revitalization of the bioinformatics-based pursuit of sequence data analysis at all levels. This renaissance can be attributed to the attractiveness of the analysis challenges introduced by large data sets and to the fact that an increasing number of compelling biological questions are now approachable with only a few experiments' worth of data, owing to the greatly increased scale and significantly lower cost of massively parallel sequencing. But whereas the data generation is straightforward, often a corresponding analytical approach to the derivation of answers is not. This fact has forged alliances between experimentalists and computational biologists as never before in genomic science, and emphasizes both the enhanced capabilities and analytical difficulties brought about by massively parallel sequencing, in contrast to the technology initially used to chart the human genome.

Defining our genomic roadmap

Variations in our sequence roadmap

Once the Human Reference Genome was in-hand¹², the efforts of the International Human Genome Project teams turned to completing all

regions of the genome to high quality¹³ on a chromosome-by-chromosome basis. Subsequently, many of these same laboratories began efforts to identify the positions of common single-nucleotide polymorphisms (SNPs) known to exist in the genome. The international SNP discovery efforts were known as the 'HapMap' projects, because they aimed to map the haplotype diversity in the human genome. These projects again required a concerted effort across many laboratories^{14–17} to characterize common SNP variation (present at 5% or greater allele frequency for the population studied) in multiple human populations.

The HapMap efforts culminated in the identification of more than 8 million common SNP positions genome-wide, most of which were generated by Sanger-based capillary methods. These were added to the dbSNP public database at NCBI (the National Center for Biotechnology Information), and so represented an important reference addendum that further emphasized the intricate genomic roadmap of individuals from various ethnic backgrounds. In addition, many HapMap SNP positions were used to increase the density of common variant sites on commercial SNP genotyping arrays, producing a research tool with which human geneticists began to evaluate large case-control cohorts for common complex disease studies. These genome-wide association studies were designed to test the 'common variants common disease' hypothesis, identifying specific loci that were associated with the occurrence of a common disease (that is, predominant in cases but not controls). Although these approaches have succeeded to various extents in identifying disease-associated SNPs¹⁸, the likely contribution that rare or 'private' SNPs make to disease susceptibility is now being investigated by combining massively parallel sequencing with case-control cohorts. One such study has shown this approach to be particularly effective in identifying rare variants found only in the genomes of affected individuals (cases) that explain the biology of the disease (in this case, extreme obesity)¹⁹.

Beyond SNPs, multiple efforts have explored the breadth of human genome diversity, demonstrating that as individuals, our roadmaps differ in many ways beyond single nucleotide differences^{20,21}. For example, there are a myriad of small- and large-scale differences in genomes, including focused insertion and deletion events (indels) genome-wide that add one to several nucleotides per event, amplification or deletion events that result in increased or decreased numbers of copies of specific genome segments (copy number polymorphisms), changes in the orientation of genome segments (inversions), and novel genome content (insertions). Although most such large-scale variations originally were observed using microarray-based methods^{22,23}, several groups have demonstrated the ability to use information from paired end or mate pair data sets towards high-resolution characterization of all classes of structural variation in the human genome^{24–26}. Structural variant discovery is achieved by examining the separation and orientation of aligned read pairs on the reference. Namely, if groups of read pairs are identified with interpair distances that are further apart or closer together than expected based on the size of inserts used in constructing the library, or with the forward and reverse reads aligned either in an unexpected orientation, or onto different chromosomes, these provide evidence for structural variation relative to the reference genome.

The 1,000 Genomes Project is adding resolution and new information to our understanding of genome diversity across all levels of variation^{27,28}. It combines the scale and cost of massively parallel human genome re-sequencing and analysis with many of the populations already studied in the HapMap projects as well as newly consented populations. To date, the Roche/454, Life Technologies SOLiD and Illumina platforms have been used for data production in this project. When the 1,000 Genomes Project is completed in 2012, information will be in hand regarding SNP, indel and structural variation for more than 2,000 individuals, and will be available through public databases to further enrich the detail of our human genome roadmap. It goes without saying that such a feat would not have been possible without the availability of massively parallel sequencing, but most will not fully appreciate the multitude of algorithmic and bioinformatic innovations required to fully mine this rich data set to its fullest.

Another area of biological enquiry that has been aided by sequencing technology is that of identifying genes affected by mutation in cancer tissue genomes. When my centre and others initially began sequencing candidate gene lists in specific cancer samples in the early 2000s, our approaches consisted of designing polymerase chain reaction (PCR) primer sets specific to the genes we thought would be mutated, amplifying those from the genomic DNA of the tumour, and sequencing with capillary-based instruments^{29–31}. The bioinformatics-based tools for identifying these mutations were largely modified from existing tools that were originally used to sequence the human genome. Although important discoveries were made during these early years in cancer genomics, the approaches were slow and expensive, and our enquiries were limited to genes whose mutations ‘made sense’ in terms of what already was known about tumour cell biology. With massively parallel sequencing, emerged the ability to pool the reaction products of thousands of PCRs and sequence them all at once, thereby reducing the cost of sequencing and dramatically increasing the rapidity with which the data could be obtained. By aligning the resulting sequence reads to the Human Genome Reference, and modifying existing algorithms for identifying mutations, mutation discovery was facilitated as well.

Shortly after these approaches were developed and published, my centre used Illumina massively parallel sequencing to sequence the complete tumour and normal genomes from a patient diagnosed with acute myeloid leukaemia (AML), and then developed methods to comparatively analyse these two genomes, identifying tumour-unique (somatic) alterations in the process³². This effort required that we develop entirely new bioinformatics-based methods to do the following: (1) ensure that we had completely sequenced the genomes to the depth and breadth needed to then identify and compare the millions of single-nucleotide differences identified in both the tumour and normal genomes (human genomes typically have about one difference per 1,000 bases when we compare any human genome data set to the reference genome sequence), and (2) ultimately, sort out the handful (typically 3,000–10,000) that are somatic, or unique to the tumour genome. Although an expensive endeavour at the time (we estimate the combination of data production and of novel bioinformatics tools development totalled \$1.6 million for the first tumour/normal pair), we and others have subsequently sequenced and analysed hundreds of human genomes using primarily Illumina and Life Technologies platforms, as the cost per genome has plummeted and the sequence data output per instrument run has increased by 100-fold. Moreover, additional algorithmic developments have combined with paired end data to reveal somatic structural variations, both for focused (small numbers of inserted or deleted nucleotides) and large events (such as inter-chromosomal translocations) and to improve our ability to find point mutations. With the newest massively parallel instruments from Illumina, the data production for each tumour and normal pair can be completed in about 8 days on a single instrument at a fully loaded cost per pair of around \$30,000. Although our analysis methods continue to be refined, the comprehensive data analysis required to characterize these paired genomes remains the most expensive and difficult aspect of whole genome re-sequencing by massively parallel methods^{33,34}.

The difficulties in the data analysis mentioned above are due to many factors, including the size and complexity of the human genome, the ever-changing read length and accuracy of next-generation sequencing data, and the computational demands needed to compute the full range of variant detection with the highest possible accuracy. In spite of refinements to these analytical methods, it is still an important and necessary step to perform orthogonal validation of discovered variants before reporting them, which further adds to the cost and time required for whole genome comparative methods.

Variations in our functional roadmap

Another benefit of these new engines is that they are allowing biologists to explore to unprecedented depths the specific differences in DNA regulation that define each tissue’s biological roadmap. In fact, massively

parallel sequencing is permitting us to answer many fundamental questions that were previously too expensive to perform at a genome-wide scale. For example, the changing associations of histones and chromosomal DNA during embryonic development, the exact placement of regulatory DNA-binding proteins on genomic DNA, the genome-wide methylation of chromosomal DNA, and the attendant alterations in gene expression levels associated with such events all can be investigated by combining an appropriate experimental front-end with a massively parallel read-out. The extent to which any or all of these measures change due to specific stimuli, over developmental time or in different tissue types also can be ascertained. One such effort aimed at performing these characterizations is the ENCODE (encyclopedia of DNA elements) Project³⁵. This project was started in 2003 and used microarray-based assays in the pilot phase, but has now moved to using massively parallel methods owing to their reduced cost, and increased resolution and speed. Such genome-wide characterization capabilities are somewhat analogous to being able to drive further and faster than ever before, while charting the geography along the routes travelled at an unprecedented level of detail. Although the scope is breathtaking, each type of experiment has a list of shared and unique considerations that the bioinformatics analysis must take into account in order to separate true signal from noise, and to deliver an accurate genomic roadmap. At the beginning of each analytical approach, these experiments all require the alignment of sequence reads onto the Human Reference Genome sequence—effectively an assignment to the chromosomal locus of origin for each DNA or RNA fragment obtained from the experiment performed.

Thus, we are using the reference in the way it was intended, as a guide to help us discover information about the human genome, and its function, regulation and alteration in health and disease. Ultimately, these experiments will enable a transformation of biomedical research and medical practice—a transformation that already has begun^{18,36}.

Charting new territories

In addition to their significant impact on our understanding of human biology, massively parallel sequencing technologies have enabled new areas of genomic enquiry that also are germane to human health. One such area is termed ‘metagenomics’—essentially, this term describes the sequencing-based characterization of DNA or RNA isolated from a mixed organism population sample obtained from its natural habitat. In human biological enquiry, metagenomic studies of the human body seek to characterize the content and complexity of microbial, viral and/or non-human eukaryotic organisms obtained from external (skin) and internal (colon, vagina) surfaces.

Depending on the body site and sampling method, variable amounts of human DNA and RNA can be simultaneously isolated, complicating the subsequent analysis to varying degrees. Although there are metagenomic studies that pre-date the availability of massively parallel sequencing instruments, the field has been transformed by rapid, inexpensive and abundant sequence data production and facile preparatory methods offered by next-generation platforms that are especially suitable for these complex genomic samples. As described earlier, the analytical challenges of these data sets have elicited an enormous amount of interest, not only in developing analysis approaches to obtain the biological information that can be mined from them, but also in applying this information to answer the fundamental questions posed by metagenomic enquiry, namely ‘what’s there?’, ‘what roles are they playing?’ and ‘how does the population change when the environment changes?’³⁷. Often, the answers to these questions lie in examining metagenomic sequences by six-frame amino-acid translation, followed by database searching with the resulting massive data set. In this regard, longer sequence reads produce longer amino acid sequences, and hence a higher probability of correctly identifying the population members and/or their metabolic capabilities. Because the Roche/454 platform has longer read lengths than either Illumina or Life Technologies SOLiD, it has been the platform of choice for metagenomics studies.

The search for unknown aetiological agents in human disease has been facilitated by massively parallel sequencing. The incredible depth

of sequencing that can be applied to a given sample, such as stool samples from an outbreak of diarrhoea, or DNA isolated from a sarcoid tumour, enables the identification of novel viruses or bacteria whose DNA or RNA will be coincidentally isolated and sequenced with that of the human samples^{38–40}. Again, the ability to detect and identify the common agent's DNA signature among the many host-derived reads requires a concerted and systematic approach to sequence data analysis. Here, longer read lengths provide more facile detection of non-human sequences and also are more straightforward to assemble, often allowing reconstruction of the aetiological agent's genome.

Finally, personal genomics—the sequencing of an individual's genome for health-related enquiry or for determining genetic predisposition to disease—emerged as an application of massively parallel sequencing once costs began to drop. James Watson, who shared the 1962 Nobel Prize in Physiology or Medicine with Francis Crick and Maurice Wilkins for discovering the chemical structure of DNA, was the first person to have his genome sequenced by massively parallel methods using the Roche/454 platform⁴¹. Other personal genomics have included solving the causative mutations in the familial Charcot-Marie-Tooth syndrome of James Lupski⁴², a noted human geneticist, and in elucidating the mutation responsible for two siblings afflicted with Miller syndrome⁴³.

Beyond these examples, there are several interesting areas to which massively parallel sequencing might be applied, effectively furthering some of the early roadmap efforts mentioned above, as well as opening interesting areas of enquiry that were previously not possible³⁶. These include (1) identifying the genomic differences between chromosomal and mitochondrial DNA derived from different tissues in a single human body, (2) establishing the gene expression profiles and patterns of all developmental and adult human tissues, (3) defining the spectrum of temporal changes in DNA methylation and histone-binding patterns of these same tissues, and (4) identifying the non-coding RNA expression profiles and their variation in human tissues.

By combining massively parallel sequencing with new whole genome DNA amplification approaches, we might also anticipate sequencing the complete genomes of single cells. Perhaps one of the most exciting possibilities addressed by this capability would permit an understanding of the genomic differences between individual tumour cells in a heterogeneous solid tumour type, such as breast cancer.

The future of sequencing

The amazing acceleration in biological enquiry enabled by the current massively parallel instrumentation is clearly just beginning. These instruments will continue to evolve, and new platforms just introduced, or under development, will have a continuing impact on biomedical research for years to come. What can we anticipate about the near-term expansion of applications using these instruments? One obvious area of expansion will be the use of massively parallel sequencing for 'genome-guided' medicine. This would involve using the speed and scope of new sequencing technologies and data analysis for diagnosis: targeted (specific genes) or whole-genome sequencing would be used to characterize the individual patient's disease, and to determine potential treatment modalities based on these data. We already have an example of this genome-guided approach; we have used whole genome sequencing and analysis to diagnose an AML patient thought to have acute promyelocytic leukaemia whose pathological diagnosis did not conform to the diagnosis obtained by cytogenetic assays, and we identified an uncommon chromosomal insertion event with a net result that mimics the common translocation (J. Welch *et al.*, manuscript submitted). Other groups have reported similar studies where massively parallel sequencing and analysis delivered answers to patients that aided their diagnosis and treatment for cancer⁴⁴ or identified the mutated gene responsible for causing a rare syndromic disease^{43,45–47}.

Certainly, one of the complications for certain types of genomic diagnoses using massively parallel instruments will be the time required to generate sequencing data (including preparatory steps and sequencing) and to completely analyse, validate and interpret these data in a medical

context. The highest capacity instruments currently available require 8–14 days to produce data. Unfortunately, these run times and the ensuing analysis may not permit the return of information in a suitable time frame, relative to the patient's need for a diagnosis.

However, interesting possible solutions to the temporal limitations of certain diagnostic applications are anticipated from the newest massively parallel systems, at present in various stages of development or in early commercial release. As these systems are capable of delivering sequencing data from single molecules of DNA as they are being sequenced—rather than as a stepwise series of nucleotide addition steps that are analysed after the sequencing instrument has finished—the time for the sequence data generation step is shortened significantly relative to next-generation systems. One such instrument from Pacific Biosciences that is being tested in early access sites monitors each one of an array of individual polymerases while DNA synthesis is occurring, in order to obtain the single molecule sequences in a minimum of 30 minutes. Other instruments in development, including those by Oxford Nanopore or IBM/Roche, use nanopore technology to identify individual DNA nucleotides as the DNA fragment passes through the nanopore, by one of several detection approaches. Although the current capacities of real-time sequencers would not permit whole human genome sequencing in a single run, the near-term application of these instruments could be on focused evaluation of specific human genes or on the genomes of aetiological agents⁴⁸ for diagnosis, prognosis or therapeutic prescription.

To be clear, although the data generation steps are relatively quick, they must be properly coupled with equally rapid sample preparation and library construction steps and with refined, time-effective sequence analysis and interpretation appropriate for the clinical setting. In this regard, single molecule systems appear to be capable of delivering read lengths that are more akin to our 2001 era capillary instruments. Indeed, once the error models for each device are well characterized, producing longer reads may return us to the level of facility in data analysis for diagnostic medicine applications that we once enjoyed while sequencing the human genome. Although this would eliminate an important bottleneck in analysis, there are multiple aspects to sort out before diagnostic sequencing becomes commonplace. Nonetheless, the future of genomic medicine via massively parallel sequencing seems imminent. As time progresses, our 'engines' continue to improve in their sophistication and power, further enabling us to explore the human genome roadmap in our continuing journey to improve human health.

1. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
3. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
4. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
5. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
6. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
7. Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
8. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).
9. Elsik, C. G. *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009).
10. Mardis, E. R. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med* **1**, 40 (2009).
11. Metzker, M. L. Sequencing technologies - the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
14. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
15. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

16. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
17. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
18. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* doi:10.1038/nature09792 (this issue).
19. Harismendy, O. *et al.* Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. *Genome Biol.* **11**, R118 (2010).
- First demonstration that rare sequence variants could be identified using next-generation sequencing in well-phenotyped cases and controls, and that functional significance in the phenotype could be assigned to the suspect variants.**
20. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
21. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
22. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
23. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
24. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
25. Chen, K. *et al.* BreakDancer: an algorithm for high resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).
26. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
27. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
- Initial structural variation data analysis resulting from 1,000 Genomes Project data, demonstrating the yield of such information from a large-scale project using next-generation sequencing.**
28. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
29. Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101**, 13306–13311 (2004).
30. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
31. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
32. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
33. Mardis, E. R. Cancer genomics identifies determinants of tumor biology. *Genome Biol.* **11**, 211 (2010).
34. Mardis, E. R. The \$1,000 genome, the \$100,000 analysis? *Genome Med* **2**, 84 (2010).
35. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
36. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* doi:10.1038/nature09764 (this issue).
37. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
38. Loh, J. *et al.* Detection of novel sequences related to African Swine Fever virus in human serum and sewage. *J. Virol.* **83**, 13019–13025 (2009).
39. Presti, R. M. *et al.* Quarantil, Johnston Atoll, and Lake Chad viruses are novel members of the family Orthomyxoviridae. *J. Virol.* **83**, 11599–11606 (2009).
40. Finkbeiner, S. R. *et al.* Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *J. Virol.* **83**, 10836–10839 (2009).
41. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
42. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
- Personal genome sequencing used to identify a rare allelic variant that causes Charcot-Marie-Tooth syndrome in the family of J. R. Lupski.**
43. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
44. Jones, S. J. *et al.* Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* **11**, R82 (2010).
45. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
46. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
- One of the first demonstrations of using exome sequencing to identify a major causative mutation in Kabuki syndrome, using genomic DNA from a small number of unrelated affected individuals.**
47. Gilissen, C. *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.* **87**, 418–423 (2010).
48. Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).

Acknowledgements I thank R. Wilson, D. Dooling and G. Weinstock for critical reading of the manuscript. J. McMichael was integral to the creation of Fig. 1.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to the author (emardis@wustl.edu).

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green¹, Mark S. Guyer¹ & National Human Genome Research Institute*

There has been much progress in genomics in the ten years since a draft sequence of the human genome was published. Opportunities for understanding health and disease are now unprecedented, as advances in genomics are harnessed to obtain robust foundational knowledge about the structure and function of the human genome and about the genetic contributions to human health and disease. Here we articulate a 2011 vision for the future of genomics research and describe the path towards an era of genomic medicine.

Since the end of the Human Genome Project (HGP) in 2003 and the publication of a reference human genome sequence^{1,2}, genomics has become a mainstay of biomedical research. The scientific community's foresight in launching this ambitious project³ is evident in the broad range of scientific advances that the HGP has enabled, as shown in Fig. 1 (see rollfold). Optimism about the potential contributions of genomics for improving human health has been fuelled by new insights about cancer^{4–7}, the molecular basis of inherited diseases (<http://www.ncbi.nlm.nih.gov/omim> and <http://www.genome.gov/GWAStudies>) and the role of structural variation in disease⁸, some of which have already led to new therapies^{9–13}. Other advances have already changed medical practice (for example, microarrays are now used for clinical detection of genomic imbalances¹⁴ and pharmacogenomic testing is routinely performed before administration of certain medications¹⁵). Together, these achievements (see accompanying paper¹⁶) document that genomics is contributing to a better understanding of human biology and to improving human health.

As it did eight years ago¹⁷, the National Human Genome Research Institute (NHGRI) has engaged the scientific community (<http://www.genome.gov/Planning>) to reflect on the key attributes of genomics (Box 1) and explore future directions and challenges for the field. These discussions have led to an updated vision that focuses on understanding human biology and the diagnosis, prevention and treatment of human disease, including consideration of the implications of those advances for society (but these discussions, intentionally, did not address the role of genomics in agriculture, energy and other areas). Like the HGP, achieving this vision is broader than what any single organization or country can achieve—realizing the full benefits of genomics will be a global effort.

This 2011 vision for genomics is organized around five domains extending from basic research to health applications (Fig. 2). It reflects the view that, over time, the most effective way to improve human health is to understand normal biology (in this case, genome biology) as a basis for understanding disease biology, which then becomes the basis for improving health. At the same time, there are other connections among these domains. Genomics offers opportunities for improving health without a thorough understanding of disease (for example, cancer therapies can be selected based on genomic profiles that identify tumour subtypes^{18,19}), and clinical discoveries can lead back to understanding disease or even basic biology.

The past decade has seen genomics contribute fundamental knowledge about biology and its perturbation in disease. Further deepening this understanding will accelerate the transition to genomic medicine (clinical care based on genomic information). But significant change rarely comes

quickly. Although genomics has already begun to improve diagnostics and treatments in a few circumstances, profound improvements in the effectiveness of healthcare cannot realistically be expected for many years (Fig. 2). Achieving such progress will depend not only on research, but also on new policies, practices and other developments. We have illustrated the kinds of achievements that can be anticipated with a few examples (Box 2) where a confluence of need and opportunities should lead to major accomplishments in genomic medicine in the coming decade. Similarly, we note three cross-cutting areas that are broadly relevant and fundamental across the entire spectrum of genomics and genomic medicine: bioinformatics and computational biology (Box 3), education and training (Box 4), and genomics and society (Box 5).

Understanding the biology of genomes

Substantial progress in understanding the structure of genomes has revealed much about the complexity of genome biology. Continued acquisition of basic knowledge about genome structure and function will be needed to illuminate further those complexities (Fig. 2). The contribution of genomics will include more comprehensive sets (catalogues) of data and new research tools, which will enhance the capabilities of all researchers to reveal fundamental principles of biology.

Comprehensive catalogues of genomic data

Comprehensive genomic catalogues have been uniquely valuable and widely used. There is a compelling need to improve existing catalogues and to generate new ones, such as complete collections of genetic variation, functional genomic elements, RNAs, proteins, and other biological molecules, for both human and model organisms.

Genomic studies of the genes and pathways associated with disease-related traits require comprehensive catalogues of genetic variation, which provide both genetic markers for association studies and variants for identifying candidate genes. Developing a detailed catalogue of variation in the human genome has been an international effort that began with The SNP Consortium²⁰ and the International HapMap Project²¹ (<http://hapmap.ncbi.nlm.nih.gov>), and is ongoing with the 1000 Genomes Project²² (<http://www.1000genomes.org>).

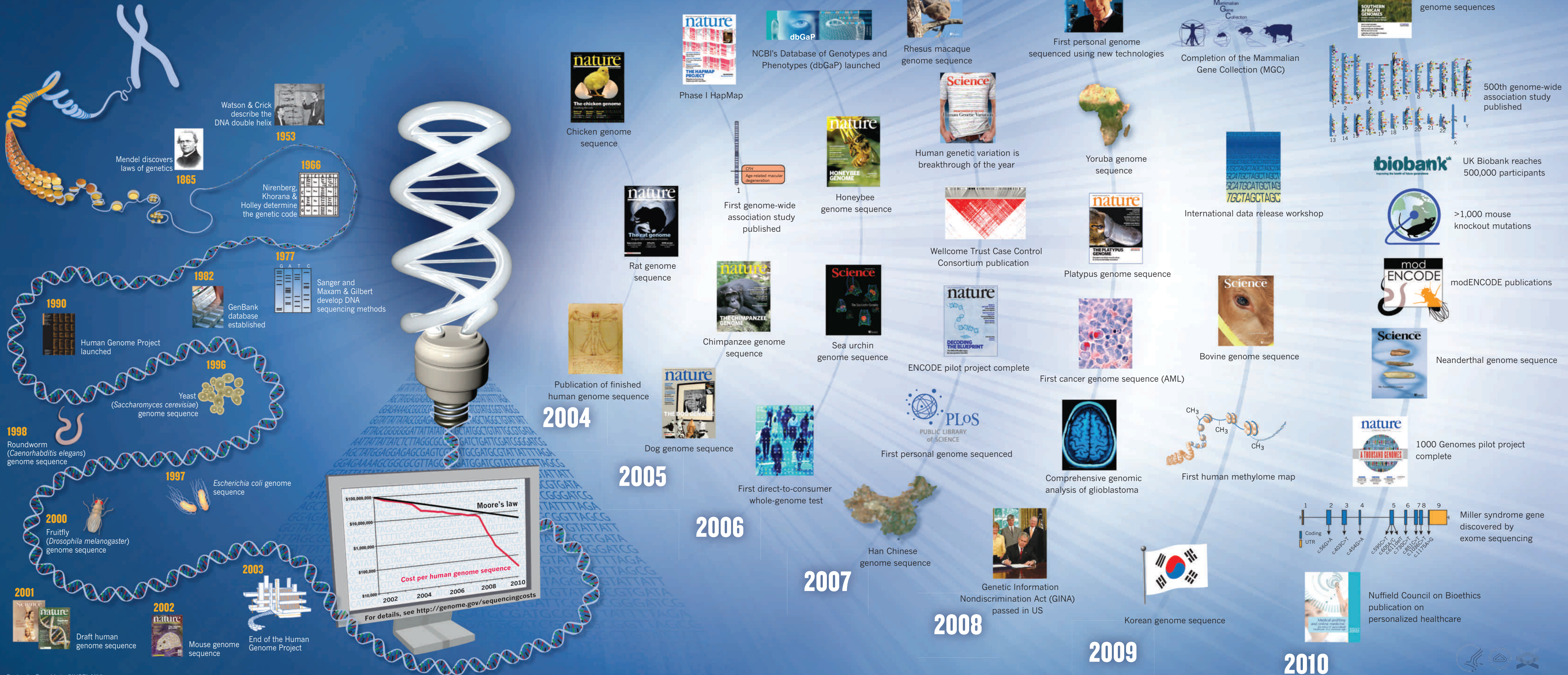
Over the past decade, these catalogues have been critical in the discovery of the specific genes for roughly 3,000 Mendelian (monogenic) diseases

Figure 1 | Genomic achievements since the Human Genome Project (see accompanying rollfold).

¹National Human Genome Research Institute, National Institutes of Health, 31 Center Dr., Bethesda, Maryland 20892-2152, USA.

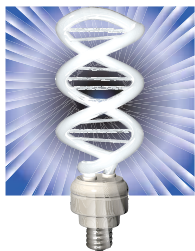
*Lists of participants and their affiliations appear at the end of the paper.

Genomic achievements since the Human Genome Project



BOX 1

The essence of genomics



Genomics grew primarily out of human genetics and molecular biology. Although the fields have much in common, genomics has several distinguishing characteristics.

Comprehensiveness. Genomics aims to generate complete data sets. Although relatively easy to define and measure for a genome sequence, attaining comprehensiveness can be

more challenging for other targets (for example, functional genomic elements or the 'proteome').

Scale. Generation of comprehensive data sets requires large-scale efforts, demanding attention to: (1) organization, often involving large interdisciplinary consortia; (2) robust data standards, to ensure high-quality data and broad utility; and (3) computational intensity (see Box 3).

Technology development. Genomics demands high-throughput, low-cost data production, and requires that resources be devoted to technology development.

Rapid data release. Large data catalogues and analytical tools are community resources. This calls for policies that maximize rapid data release (harmonized internationally), while respecting the interests of the researchers generating the data and the human participants involved in that research^{99,100}.

Social and ethical implications. Genomics research and the many ways in which genomic data are used have numerous societal implications that demand careful attention (Box 5).

(<http://www.ncbi.nlm.nih.gov/omim>) and in establishing genetic associations between more than 900 genomic loci and complex (multigenic) traits, many of them diseases (<http://www.genome.gov/GWASstudies>). New genes and pathways have been implicated in disease, unexpected genetic connections among diseases have been identified, and the importance of non-coding variants in human disease has been highlighted. Together, these findings have accounted for a portion, but not all, of the heritability for many complex diseases²³. Complete characterization of the genetics of complex diseases will require the identification of the full spectrum of human genomic variation in large, diverse sample sets.

Comprehensive catalogues of genetic variation in non-human species are similarly valuable. For example, understanding genetic variation in insect disease vectors may help inform the development of new strategies to prevent disease transmission, whereas knowledge about variation among microbial pathogens may lead to more robust vaccine-design strategies and novel therapeutics.

Catalogues of functional elements in the human genome, and the genomes of other species, are also being developed ('functional elements' include genes that encode proteins and non-coding RNAs; transcripts, including alternative versions; protein–nucleic-acid interaction sites; and epigenomic modifications). The ENCYclopedia Of DNA Elements (ENCODE)²⁴ (<http://genome.gov/encode>) and modENCODE (<http://genome.gov/modencode>) projects are developing catalogues of functional elements in the human genome and in the genomes of *Caenorhabditis elegans*²⁵ and *Drosophila melanogaster*²⁶, respectively. But building a truly comprehensive catalogue of functional elements for any multicellular organism will require analysis of a large number of biological samples using many assays. Novel high-throughput, cost-effective technologies, and new reagents (see below), are needed to complete the human, fly and worm catalogues and compile catalogues of other genomes (for example, mouse and rat).

Biomedical research would benefit immensely from the availability of additional catalogues, for example of DNA modifications (epigenomics),

gene products such as RNAs (transcriptomics) and proteins (proteomics), and indirect products of the genome such as metabolites (metabolomics) and carbohydrates (glycomics). Undertaking such large efforts will depend on both demand and the opportunity to cost-effectively assemble data sets of higher quality and greater comprehensiveness than would otherwise emerge from the combined output of individual research projects. Although the generation of some of these catalogues has already begun, major advances in technologies and data analysis methods are needed to generate, for example, truly comprehensive proteomic data sets and resources.

Additional insights will come from combining the information from different catalogues. For example, analysing genetic variation within functional elements will be particularly important for identifying such elements in non-coding regions of the genome. To this end, the GTEx (Genotype-Tissue Expression) project (<http://www.commonfund.nih.gov/GTEx>) has been established to map all sites in the human genome where sequence variation quantitatively affects gene expression.

New tools for genomics research

Technology development has driven genomics. Both revolutionary (new methods, reagents and instruments) and evolutionary (incremental improvement in efficiency and output) technology development have been critical for achieving the remarkable increases in throughput and reductions in costs of DNA sequencing and other genomic methods. However, the inherent complexity of biology means that current technology is still not adequate for obtaining and interpreting the next generation of genomic data. Technological challenges include the design, synthesis and use of synthetic DNAs, and the measurement of cell- and organism-level phenotypes. Orders-of-magnitude improvements in throughput, cost-effectiveness, accuracy, sensitivity and selectivity of genomic technologies will require novel approaches^{27,28}.

Massively parallel DNA sequencing²⁹ has enabled a three-to-four orders-of-magnitude fall in the cost of genome sequencing (Fig. 1; see accompanying paper³⁰ and <http://genome.gov/sequencingcosts>). Nevertheless, sequencing a whole human genome remains much too expensive for most human disease studies, each of which can involve thousands or tens of thousands of individuals. Even in the case of well-understood coding regions (exons), sequencing errors complicate downstream analyses, and current sequencing error rates hinder reliable analysis of the remaining, poorly understood 98% of the genome. Perhaps most importantly, very low cost and extremely high accuracy will be critical for the routine clinical use of genome sequencing (for example, genetic screening of newborn babies^{31,32}).

Structurally complex genomic regions, which are known to have a role in human disease⁸, remain inherently difficult to sequence, even with the new DNA sequencing technologies. Additional technological improvements (for example, much longer read lengths) are needed to sequence such complex regions and to finish any specific region efficiently. Only with the ability to sequence entire genomes at very high accuracy, completeness and throughput will genome sequencing reach its full potential.

Some clinical applications (for example, rapid genomic analysis of tumours or microbiomes) may benefit from complete genomic sequencing in hours rather than weeks ('Making genomics-based diagnostics routine', Box 2). Although speed may be less important for research applications, it could have profound benefits in certain situations in the clinic. As genomics permeates clinical practice, point-of-care implementations will be needed, including in locations with minimal infrastructure. Separate technologies are likely to emerge for the research and clinical settings.

Analysis of functional genomic elements will require high-specificity affinity reagents (for example, antibodies or other tagging molecules) for all transcription factors, nucleic-acid-binding proteins, histone forms and chromatin modifications. These reagents must function well in a number of assays to be maximally useful. Several large-scale efforts to generate such reagents are under way^{33,34} (see also <http://www.commonfund>).

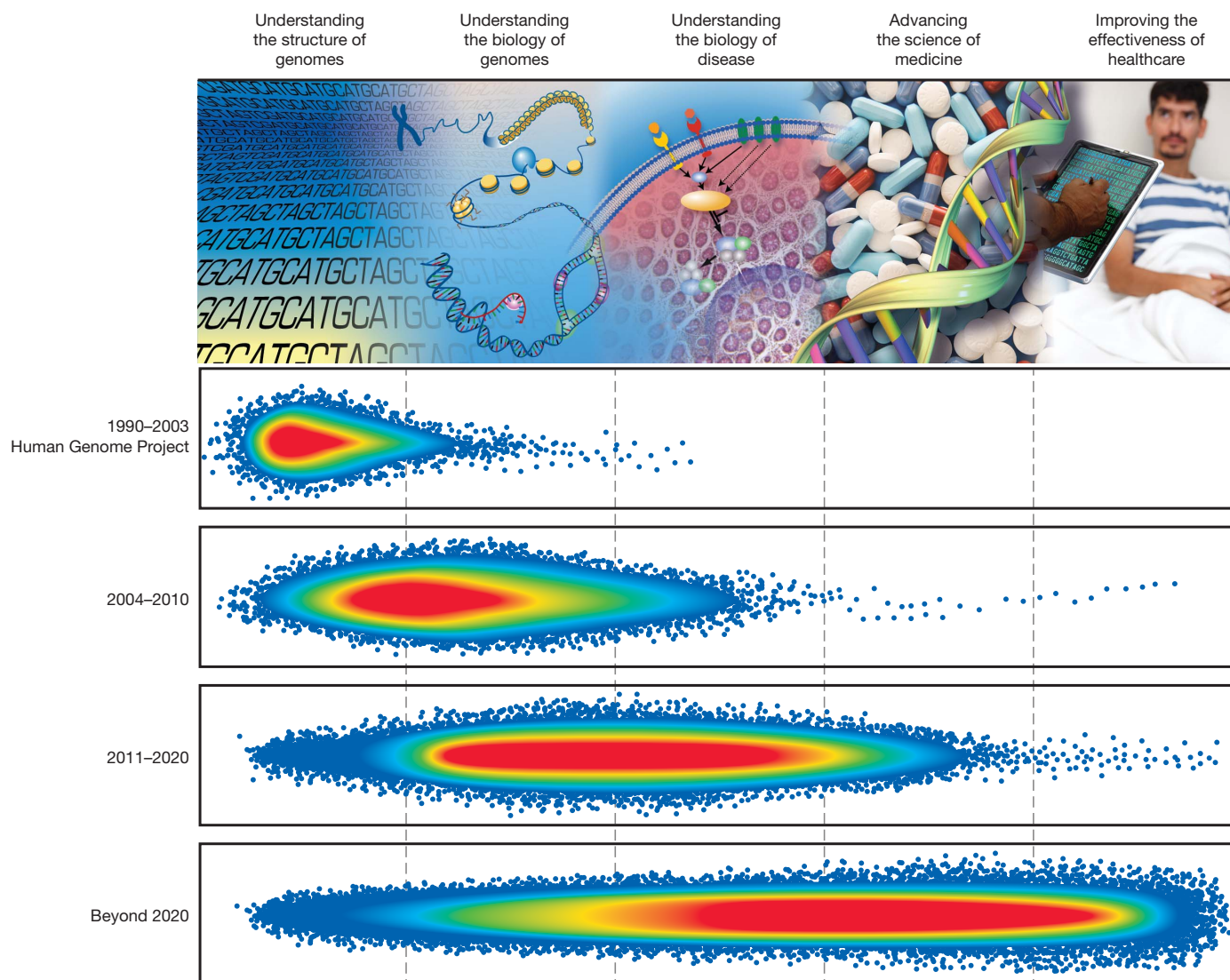


Figure 2 | Schematic representation of accomplishments across five domains of genomics research. The progression from base pairs to bedside is depicted in five sequential, overlapping domains (indicated along the top). Genomic accomplishments across the domains are portrayed by hypothetical, highly schematized density plots (each blue dot reflecting a single research

accomplishment, with green, yellow and red areas reflecting sequentially higher densities of accomplishments). Separate plots are shown for four time intervals: the HGP; the period covered by the 2003 NHGRI vision for the future of genomics research¹⁷; the period described here (2011–2020); and the open-ended future beyond 2020.

nih.gov/proteincapture and <http://antibodies.cancer.gov>), but current approaches will probably not produce the full spectrum of reagents of the required specificity and utility. Suitable affinity reagents for larger-scale proteomic analyses pose an even greater challenge.

Most assays of functional genomic elements are currently limited by the need for a large number of cells, so many experiments are now performed with either tissue culture cells (which may not accurately reflect *in vivo* states) or heterogeneous tissue samples (in which sub-tissue-specific patterns may go undetected). Developing methods for producing accurate cell-specific profiles of single cells is a challenge. Analysing genomic data requires integration of multiple data types (Box 3). Robust analysis of promoters, for example, typically involves integration of data on transcription factor binding, protein complex formation, transcription start sites and DNase hypersensitive sites. Improved data integration approaches will require new algorithms and robust computational tools (Box 3).

The spatially and temporally dynamic nature of genomic regulation (see below) presents another formidable challenge to the comprehensive identification of all functional elements, as some critical regulatory processes only occur during brief developmental periods or in difficult-to-access

tissues. New methods for *in situ* and real-time analysis will be necessary to understand fully the choreography of gene regulation.

Phenotypes arise from complex interactions among genes, cells, tissues, organs and the environment. Ultimately, the ability to co-analyse variation and phenotypic data will be critical for generating reliable inferences about disease-causing loci, genotype–phenotype correlations, and both gene–gene and gene–environment interactions. Therefore, better technologies for measuring phenotypes, behaviours, exposures and other environmental variables will be required.

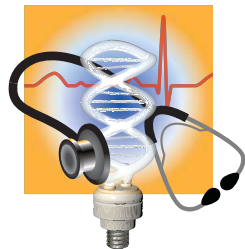
Understanding fundamental principles of biology

Comprehensive genomic catalogues are the ‘parts lists’, and just as such a list is not sufficient to understand how a machine functions, genomic catalogues are not sufficient to understand biological processes.

Recent advances have greatly improved our understanding of the importance of non-coding regions in the human genome for gene regulation, chromosome function and the generation of untranslated RNAs³⁵. This is further supported by the finding that the great majority of trait-associated regions identified by genome-wide association studies fall in non-coding sequences³⁶. New technologies, experimental strategies and

BOX 2

Imperatives for genomic medicine



Opportunities for genomic medicine will come from simultaneously acquiring foundational knowledge of genome function, insights into disease biology and powerful genomic tools. The following imperatives will capitalize on these opportunities in the coming decade.

Making genomics-based diagnostics routine. Genomic technology

development so far has been driven by the research market. In the next decade, technology advances could enable a clinician to acquire a complete genomic diagnostic panel (including genomic, epigenomic, transcriptomic and microbiomic analyses) as routinely as a blood chemistry panel.

Defining the genetic components of disease. All diseases involve a genetic component. Genome sequencing could be used to determine the genetic variation underlying the full spectrum of diseases, from rare Mendelian to common complex disorders, through the study of upwards of a million patients; efforts should begin now to organize the necessary sample collections.

Comprehensive characterization of cancer genomes. A comprehensive genomic view of all cancers^{4–7} will reveal molecular taxonomies and altered pathways for each cancer subtype. Such information should lead to more robust diagnostic and therapeutic strategies and a roadmap for developing new treatments^{74,75}.

Practical systems for clinical genomic informatics. Thousands of genomic variants associated with disease risk and treatment response are known, and many more will be discovered. New models for capturing and displaying these variants and their phenotypic consequences should be developed and incorporated into practical systems that make information available to patients and their healthcare providers, so that they can interpret and reinterpret the data as knowledge evolves.

The role of the human microbiome in health and disease. Many diseases are influenced by the microbial communities that inhabit our bodies (the microbiome)¹⁰¹. Recent initiatives^{102,103} (<http://www.human-microbiome.org>) are using new sequencing technologies to catalogue the resident microflora at distinct body sites, and studying correlations between specific diseases and the composition of the microbiome¹⁰⁴. More extensive studies are needed to build on these first revelations and to investigate approaches for manipulating the microbiome as a new therapeutic approach.

computational approaches are needed to understand the functional role of non-coding sequences in health and disease.

Genomics will also contribute new technologies and resources to the analysis of gene-interaction networks. Network analysis will benefit from understanding the dynamics of gene expression, protein localization and modification, as well as protein–protein and protein–nucleic-acid associations. The ultimate challenge will be to decipher the ways that networked genes produce phenotypes. Genomics can contribute to solving this problem by providing data from systematic large-scale studies of gene expression that include determination of cellular responses to genetic changes, external perturbations and disease (see <http://www.commonfund.nih.gov/LINCS>). Here too, robust new computational tools are needed specifically to access, analyse and integrate large, complex data sets, and to develop predictive models, new visualization technologies and a 'knowledge base' of networks (Box 3).

Ultimately, human biology must be understood in the context of evolution. Comparative genomic studies have revealed the most highly conserved (and probably functional) portions of human, mammalian

and vertebrate genomes³⁷. Evolutionary relationships also underlie the use of model organisms in functional studies, and diverse data sets from unicellular organisms to mammals³⁸ will lead to key insights about genome function and biological pathways.

Despite their necessity, however, large-scale genomic studies alone will not be sufficient for gaining a fundamental understanding of biology. Most of the data analysis and interpretation will actually come from individual research efforts. Indeed, a primary motivation for the development of genomics (and other '-omics' disciplines) has been to generate data catalogues and technological tools that empower individual investigators to pursue more effective hypothesis-driven research.

Understanding the biology of disease

All diseases are influenced by genetic variation (inherited and/or somatic), environmental agents and/or health behaviours, and there is increasing evidence for epigenomic contributions^{39,40}. Using genomics is essential to understand both the normal and disease-related functions of the genetic and epigenetic⁴¹ contributors to disease, and the cellular pathways and biological processes in which they are involved, an understanding that is critical to the development of improved strategies for diagnosis, prevention and therapeutic intervention.

The power of genomic approaches to elucidate the biology of disease is illustrated by the study of Crohn's disease. A decade ago, the mechanisms underlying this debilitating gastrointestinal disorder were opaque. Since then, genome-wide association studies have identified dozens of genomic regions harbouring genetic variants conferring risk for Crohn's disease⁴². Analyses of genes in these regions have revealed key, previously unappreciated roles in the disease for several physiological processes, including innate immunity, autophagy and interleukin (IL)-23R signalling^{43,44}. Cellular models have been developed and used both to document the pathogenicity of specific mutations and to extend knowledge of the relevant biological pathways⁴². Chemical screens have been designed^{42,45} to identify new candidate therapeutic agents. Furthermore, animal models have been developed that accurately model the effects of causal variants found in patients. In sum, the use of genomic approaches to identify risk-conferring variants has catalysed molecular, cell biological and animal model studies that have led to a better understanding of Crohn's disease and the development of novel therapies. This and other examples^{11,46,47} justify the optimism about genomics' potential to accelerate the understanding of disease.

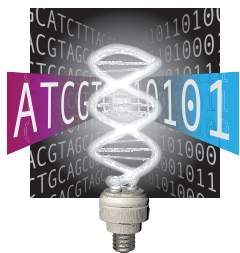
Genetic and non-genetic bases of disease

Genomics will allow the compilation of rich catalogues spanning the full spectrum of germline variants (both common and rare) conferring risk for inherited disease ('Defining the genetic components of disease', Box 2). Catalogues of somatic mutations that contribute to all aspects of tumour biology for each major cancer type are under development^{44,48} ('Comprehensive characterization of cancer genomes', Box 2; see also <http://www.icgc.org> and <http://www.sanger.ac.uk/perl/genetics/CGP/cosmic>). Effective partnerships among investigators with genomics expertise, those with in-depth knowledge of specific diseases, and patients will lead to the definition of the pathways from genetic variant to disease. Success will require improved definitions and measurements of phenotypes, new databases ('Practical systems for clinical genomic informatics', Box 2) and novel experimental strategies, such as studies of individuals associated with extremes of risk and phenotype and studies of risk-reducing variants (which may provide guides to new therapeutics and approaches to disease management).

Genome-wide association studies have implicated hundreds of non-coding genomic regions in the pathogenesis of complex diseases⁴⁹, creating a major challenge. Establishing disease causality of non-coding variants will be considerably more difficult than identifying causal variants in protein-coding sequences. In developing methods to characterize the functional landscape of non-coding DNA (discussed above), particular attention must be paid to establishing novel strategies for identifying non-coding variants that influence disease. Actually, disease research

BOX 3

Bioinformatics and computational biology



The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges.

Data analysis. Computational tools are quickly becoming inadequate for analysing the amount of genomic data

that can now be generated, and this mismatch will worsen. Innovative approaches to analysis, involving close coupling with data production, are essential.

Data integration. Genomics projects increasingly produce disparate data types (for example, molecular, phenotypic, environmental and clinical), so computational approaches must not only keep pace with the volume of genomic data, but also their complexity. New integrative methods for analysis and for building predictive models are needed.

Visualization. In the past, visualizing genomic data involved indexing to the one-dimensional representation of a genome. New visualization tools will need to accommodate the multidimensional data from studies of molecular phenotypes in different cells and tissues, physiological states and developmental time. Such tools must also incorporate non-molecular data, such as phenotypes and environmental exposures. The new tools will need to accommodate the scale of the data to deliver information rapidly and efficiently.

Computational tools and infrastructure. Generally applicable tools are needed in the form of robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists. Adequate computational infrastructure is also needed, including sufficient storage and processing capacity to accommodate and analyse large, complex data sets (including metadata) deposited in stable and accessible repositories, and to provide consolidated views of many data types, all within a framework that addresses privacy concerns. Ideally, multiple solutions should be developed¹⁰⁵.

Training. Meeting the computational challenges for genomics requires scientists with expertise in biology as well as in informatics, computer science, mathematics, statistics and/or engineering. A new generation of investigators who are proficient in two or more of these fields must be trained and supported.

may have a leading role in illuminating the fundamental biology of non-coding sequence variation and its phenotypic implications.

A full understanding of disease will require capturing much of the genetic variation across the human population⁵⁰. Accomplishing this will involve collaborations with relevant communities, taking into account how genomics is understood and perceived by different racial, ethnic and cultural groups, to form effective partnerships that will ensure that such research is sound and ethically conducted. Given the history of incidents leading to misunderstanding and mistrust⁵¹, this is an area ripe for innovative approaches.

A complete understanding of disease also requires the annotation and correlation of genomic information with high-quality phenotypic data. Obtaining phenotypic data that are both thorough and accurate enough to be analysed in conjunction with high-quality genomic and environmental data requires meticulous application of phenotyping methods, improved definitions of phenotypes, new technologies, and the consistent use of data standards⁵² (<http://www.phenx.org>). To interrogate this information effectively, widely accessible databases containing extensive phenotypic information linked to genome sequence data (genotype) are

needed⁵³. Such efforts will benefit greatly from the linkage of genomic information to data gathered in the course of actual clinical care, such as in electronic medical/health records. Research is needed to help formulate evidence-based solutions for the complex ethical, legal and regulatory challenges associated with generating and using such linkages.

The integration of genomic information and environmental exposure data can help to understand the links between biological factors and extrinsic triggers, providing a much fuller understanding of disease aetiology. Obtaining such integrated data sets can be immeasurably aided by large-scale prospective cohort studies, which allow robust analyses of genetic and environmental risks across the human lifespan, but present unique challenges in scale-up and implementation⁵⁴. Several such cohort studies have been initiated (<http://www.p3g.org/secretariat/memb.shtml> and <http://www.nationalchildrensstudy.org>) or proposed⁵⁵.

Studies of non-human organisms can help to characterize disease-implicated variants and understand their biology, providing valuable insights about health and disease. Genomics has enhanced the utility of both widely used models (for example, yeast, fruitflies, worms, zebrafish, mice and rats) and less commonly used organisms that provide good models for human disease (for example, the ferret for studying influenza, the armadillo for leprosy, and the prairie vole for social behaviour, including autism). New animal models developed on the basis of genomic insights are enormously valuable and should be made broadly available. A particularly interesting application of genomics involves microbes. The biological relevance of human-microbe interactions is both obvious (in infectious diseases) and relatively unexplored (in the maintenance of human health). Advances in DNA sequencing technologies and new approaches for data analysis have contributed to the emergence of metagenomics, which offers unprecedented opportunities for understanding the role of endogenous microbes and microbial communities in human health and disease ('The role of the human microbiome in health and disease', Box 2).

Human participants in genomics research

Effective genomics research needs continual, broad and representative public participation, and depends on developing trust and informed partnerships between researchers and different segments of society. In both genomics research and medicine, it is particularly important to recognize the need for balance among a range of competing considerations (Box 5). And as in all biomedical research, it is imperative to recognize and respect the distinctions between research and clinical care.

The oversight system for human subjects' protection is based on principles related to identifiability, risk-benefit assessment, equitable selection of participants and considerations of informed consent. However, genomics research can sometimes challenge our ability to apply these principles. For example, existing definitions of identifiability are problematic because even modest amounts of genomic sequence are potentially identifying and refractory to anonymization. Other types of genomic information (for example, transcript and microbiome profiles) may also be identifying. In addition, concepts of genomic privacy vary among individuals and cultures.

Genomics research challenges standard approaches to informed consent because it is necessary to design consent language that fully accounts for the broad utility that genomic data can offer beyond the immediate study. Such challenges are magnified in large studies that involve many thousands of participants. Studies that use archived samples pose distinct problems because such samples were often collected using consent processes that did not anticipate the potential identifiability of genomic data or the value of broad, long-term data sharing.

In consideration of the unique and potentially sensitive nature of genomic information, the framework for oversight of genomics research involving human subjects should be re-examined to ensure appropriate protections of all participants. Although legal protections to prevent inappropriate use of genetic information have been developed in some countries^{56,57}, best practices for informed consent processes and improved policies on the use of existing samples and data are needed⁵⁸.

BOX 4

Education and training



Realizing the benefits of genomics will require an educated public who can understand the implications of genomics for their healthcare and evaluate the relevant public policy issues. Clinical professionals will need to be trained to work within interdisciplinary teams. The development of effective education

and training efforts will require that diverse communities be engaged, so that all can appropriately benefit.

Strengthening primary and secondary education. If general science literacy is to improve, including an understanding of probability and risk that is relevant to genomic medicine, biological sciences curricula during primary and secondary education need to change. This, in turn, requires improvements in the training of science educators.

Conducting public outreach. Education programmes are needed to promote lifelong public understanding and awareness of the role of genomics in human health and other areas.

Building healthcare providers' genomic competencies. All healthcare providers must acquire competency in genomics to provide services appropriate for their scope of practice. Genomics needs to be better integrated into the curricula of healthcare professional education programmes, as well as their licensing and accrediting processes.

Preparing the next generation of genomics researchers. Many disciplines beyond bioinformatics/computational biology and medicine, including mathematics, public health, engineering and the humanities, have relevance to genomics and its uptake. The number of trainees acquiring expertise in both genomics and one or more related fields must increase. The diversity of the genomics workforce must also expand.

Another acute challenge arises from the fact that genomics research inevitably reveals information about participants' risk factors or disease status for disorders and traits not being directly studied (so-called incidental findings). Additional research and policies are needed to guide decisions about whether, when, and how to return individual research findings (especially incidental findings) to research participants^{59–61}. Guidance is also needed to account for the likelihood that the interpretation of genomic information will evolve over time.

Identifiability, privacy, informed consent and return of results are not the only issues pertaining to research participants that are raised by genomics. Research is also needed to understand issues related to ownership of samples and data, data access and use, intellectual property, and benefit sharing, among others.

Advancing the science of medicine

The science of medicine and the practice of medicine (that is, the provision of healthcare) are distinct domains. Our burgeoning knowledge of the human genome is beginning to transform the former, and there are already examples where genomic information is now part of the standard of care^{62–64}. Genomic discoveries will increasingly advance the science of medicine in the coming decades (Fig. 2), as important advances are made in developing improved diagnostics, more effective therapeutic strategies, an evidence-based approach for demonstrating clinical efficacy, and better decision-making tools for patients and providers. Realistically, however, a substantial amount of research is usually needed to bring a genomic discovery to the bedside, as initial findings indicating potential benefits must be followed by clinical studies to demonstrate efficacy and effectiveness⁶⁵.

Diagnostics

Over the next decade, the variant genes responsible for most Mendelian disorders will be identified and, for some number, such knowledge will lead to the development of practical treatments. A more immediate benefit will be an accurate diagnosis that, even in the absence of a treatment, can be clinically valuable. A rapid, accurate diagnosis cuts short the 'diagnostic odyssey' that often involves many false leads and ineffective treatments, can reduce healthcare costs, and provide psychological benefit to patients and families.

Beyond Mendelian disorders, a major benefit of genomic (and other 'omic') information will come from accurate subclassification of diseases. As shown for breast cancer¹⁹, understanding the 'molecular taxonomy' of a disease can help distinguish different conditions that have common pathophysiological or morphological features, yet respond to different treatments.

Therapeutics

Genomic information can be used in many ways for developing improved therapeutics. The following discussion focuses on pharmaceuticals, where genomic information can inform target identification, rational drug design, genomics-based stratification in clinical trials, higher efficacy and fewer adverse events from genotype-guided drug prescription (pharmacogenomics), as well as guide the development of gene therapy strategies. Genomic information will also inform therapeutic approaches based on dietary, behavioural and lifestyle interventions, modification of environmental exposures, and other population-based or societal interventions that have genotype-specific effects^{66–68}.

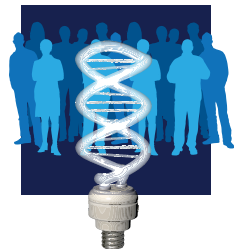
The systematic development of a pharmaceutical requires the discovery and validation of a disease-relevant target in the relevant cells. Traditionally, targets have been identified biochemically, one at a time. The more thorough understanding of disease potentiated by genomics will bring extraordinary opportunities for identifying new targets for drug development⁶⁹. Using detailed information about a disease, candidate therapeutic agents (for example, small molecules, antibodies and other proteins, and small interfering RNAs) can be identified by high-throughput screening methodologies or developed by molecular design technologies. It must be noted, however, that many of the subsequent steps in drug development (for example, medicinal chemistry, pharmacokinetics and formulation) do not involve genomics, and cannot be expected to be improved by it. The development of new pharmaceuticals based on genomic knowledge of specific targets and their role in disease has already been markedly successful^{70–73}, and is becoming increasingly commonplace, particularly for cancer drug development^{74,75}.

At the same time, understanding the underlying disease biology based on genomic information does not guarantee new therapeutics. For example, although some human disease genes (such as those for sickle cell anaemia, Huntington's disease, and cystic fibrosis) were identified more than two decades ago, the development of suitable therapies for these disorders has been much slower than anticipated. Although there have been recent promising developments^{13,76}, success is by no means certain in all cases.

Another significant opportunity offered by genomics is improved design of clinical trials⁷⁷. Currently, many clinical trials treat the tested population as genetically homogeneous. But stratification of trial participants using genomic information can allow the use of smaller numbers of participants and increase statistical power for establishing effectiveness and reducing morbidity. An example is gefitinib, for which survival benefit was only documented by analysis in a genomically selected population⁷⁸. Genomics should also allow the identification of individuals genetically susceptible to adverse reactions⁷⁹. Correlation of genomic signatures with therapeutic response will enable the targeting of appropriate patients at appropriate stages of their illness in clinical trials, resulting in more effective drugs as well as better dosing and monitoring. It will also significantly affect the information provided to prospective research participants regarding the potential for medical benefit directly related to trial participation, a topic of intense controversy for early-phase clinical trials⁸⁰.

BOX 5

Genomics and society



Effectively examining the societal implications of genomic advances requires collaborations involving individuals with expertise in genomics and clinical medicine and experts in bioethics, psychology, sociology, anthropology, history, philosophy, law, economics, health services research and related disciplines.

Psychosocial and ethical issues in genomics research. These include ensuring appropriate protection of human research participants and addressing the perceptions of risks and benefits of participating in genomic studies; expanding the diversity of research cohorts; incorporating biological ancestry markers and self-identified race and ethnicity as variables in genomic studies; accomplishing effective community engagement; and including vulnerable populations (for example, children and the disabled) and deceased individuals in genomics research.

Psychosocial and ethical issues in genomic medicine. These include communicating with patients about the uncertainty and evolving nature of predictions based on genomic information; interpreting information from direct-to-consumer genetic tests; ensuring fair access to genomic medicine; assessing the effectiveness of genomically informed diagnostics and therapeutics; using genomic information to improve behaviour change interventions; addressing issues associated with pre-implantation, prenatal and postnatal genetic diagnoses; and determining how constructs of race and ethnicity relate to the biology of disease and the potential to advance genomic medicine.

Legal and public policy issues. These include intellectual property in genomics; insurance reimbursement for genomic services; regulation of genetic testing; regulatory and non-regulatory approaches for dealing with direct-to-consumer genetic testing; the regulation of pharmacogenomics and genomics-based therapeutics; protection against genetic discrimination and stigmatization; and uses of genomics in non-medical settings.

Broader societal issues. These include the implications of increasing genomic knowledge for conceptualizing health and disease; for understanding identity at the individual and group levels, including race and ethnicity; for gaining insights about human origins; and for considering genetic determinism, free will and individual responsibility.

Pharmacogenomics is another direct clinical application of genomic medicine. Genetically guided prescription of the antiretroviral drug abacavir is now the standard of care for HIV-infected patients⁸¹, and it is likely that the use of tamoxifen⁸², clopidogrel⁸³ and possibly warfarin⁸⁴ will soon benefit from genetic considerations. Realistically, however, pharmacogenomics will not be useful for all drugs, such as those for which metabolism is not affected by genetic variation or for which there are redundant metabolic pathways. As in any other area of medicine, actual patient benefit must be demonstrated before routine clinical use of a pharmacogenomic test⁶⁵.

An evidence base for genomic medicine

The effectiveness of genomic information in tailoring interventions and, ultimately, improving health outcomes must be demonstrated. Genomically informed interventions (for example, pharmacogenomic tests or the use of genomics-based information to change risk behaviour) must be evaluated with a portfolio of research approaches, including retrospective analyses, prospective studies, clinical trials and comparative effectiveness studies, to evaluate their impact on decision making,

health outcomes and cost. This will also help to avoid harm to patients or the wasting of time and resources⁶⁸. However, although a substantial evidence base before clinical introduction is ideal, there can be costs in delaying the implementation of useful genomics-based strategies. In some situations, genomic information may provide opportunities to develop and use innovative clinical trial designs that lead to provisional approval with continued study. Informed and nuanced policies for healthcare payer coverage could also facilitate provisional implementation while definitive data are accrued.

Genomic information and the reduction of health disparities

Most documented causes of health disparities are not genetic, but are due to poor living conditions and limited access to healthcare. The field of genomics has been appropriately cautioned not to overemphasize genetics as a major explanatory factor in health disparities⁸⁵. However, genomics research may still have a role in informing the understanding of population differences in disease distribution, treatment response and the influence of gene–environment interaction and epigenomics on disease and health^{86,87}. For example, a few genetic variants can be correlated with population differences associated with an increased risk for several diseases with documented prevalence disparities, such as prostate cancer⁸⁸ and kidney disease⁸⁹. Although the results of most genomic studies will apply broadly, it is important to identify any specific genetic factors that may be associated with disparate disease risk, incidence, or severity among population groups.

Barriers to obtaining the benefits of genomics need to be identified and addressed. It will be important to recognize and understand how genomics researchers and research participants conceptualize and characterize human groups and whether or how such categorizations shape research outcomes. Many group-based social identities, most notably those reflecting race, ethnicity and nationality, include ancestry and morphology as bases of categorization⁹⁰. When analysing phenotypic data, innovative approaches will be needed to tease apart the many confounders that co-vary with social identity. Progress in parsing the interactions among multiple genetic, environmental and social factors promises to provide more accurate predictions of disease risk and treatment response. Most importantly, as genomics continues to be applied in global healthcare settings, it must not be mistakenly used to divert attention and resources from the many non-genetic factors that contribute to health disparities, which would paradoxically exacerbate the problem.

Delivering genomic information to patients

The routine use of genomics for disease prevention, diagnosis and treatment will require a better understanding of how individuals and their healthcare providers assimilate and use such information. The amount and heterogeneous nature of the data, which will include both expected and unexpected results, will antique current mechanisms for delivering medical information to patients.

Healthcare professionals will need to be able to interpret genomic data, including those from direct-to-consumer services, that are relevant to their scope of practice and to convey genetic risk to their patients. Patients will need to be able to understand the information being provided to them and to use that information to make decisions. Implementation research will help define the best ways to convey the uncertainties and complexities of genomics-based risk information to individuals and their families, how such information is understood, and how it influences health-related behaviour. Principles should be developed for guiding decisions about acquiring genomic information. These principles will have to balance the potential benefits of new preventive measures and therapeutics with economic impact and the potential for harm.

Achieving effective information flow will require an understanding of the issues related to achieving genomic medicine literacy by healthcare providers and consumers (Box 4) and the influence of genomic information on an array of health behaviours⁶⁸. Additional research should

investigate the impact of various factors (for example, family history and underlying motivations) on patients' ability to reduce their risk. Here too, evidence-based best practices are needed to ensure that patients have adequate information, access to appropriate healthcare services, and suitable follow-up to help them use their genomic information. These best practices should also inform the development and implementation of evidence-driven regulatory policies that enhance the public benefit of genomics, but at the same time protect the public from inaccurate claims and the dissemination of unreliable information.

Additional challenges will arise as genomics becomes part of global medicine. Strategies that take into account differences in healthcare practices and systems will be required to realize the potential of genomics to prevent and treat disease around the world.

Improving the effectiveness of healthcare

Clinical deployment of genomics has already begun in a small number of cases; widespread implementation, however, will take many years (Fig. 2) and must be an iterative process that continually incorporates new findings. To obtain the healthcare benefits of genomics, various important issues need to be considered.

Electronic medical/health records

Viable electronic medical/health records systems capable of handling family history and genomic data are required to fully utilize genomic information for patient care. Existing clinical informatics architectures are largely incapable of storing genome sequence data in a way that allows the information to be searched, annotated and shared across healthcare systems over an individual's lifespan. Innovative approaches are needed to assimilate a patient's genomic information^{91,92}, as are user-friendly systems that permit retrieval and queries by healthcare providers⁹³. There are intensive efforts to create new technologies and systems that bring the electronic medical/health record into routine use⁹⁴. The value of such records for genomics research has been demonstrated^{95,96}. In developing these systems, close attention to the ethical, legal and regulatory complexities is essential. Public concern about health information privacy is already widespread. Although the concern may be greater for genomic information, it is inherent to medical information and can be addressed⁹⁷ through the interaction of genomics experts with the medical informatics and policy communities.

Demonstrating effectiveness

Demonstrating utility will be critical for the widespread adoption of genomic medicine, including reimbursement for services. The thresholds for evidence of benefit and harm vary across stakeholders, and defining robust metrics for measuring utility is an important research objective. Such studies will need to assess patient outcomes (including morbidity and mortality or, minimally, widely accepted surrogate health markers).

The effective uptake of genomic medicine will require productive interactions with the regulatory systems in each country. Addressing these and other rapidly emerging issues will require sustained, yet agile, collaborative efforts by the research, regulatory and healthcare communities, as well as new research models that involve rapid iterative cycles. Rather than using traditional clinical trials, such an approach could involve practice-based interventions spanning the range of clinical, patient-reported and economic outcomes measured at the level of individuals, practices and systems.

Educating healthcare professionals, patients and the public

Education at many levels will be critical for the successful introduction of genomics into healthcare (Box 4). Genomics-based healthcare is no different from standard healthcare in being a combined responsibility of the patient and medical professionals, and all must be well informed. As genomics moves into routine clinical practice, innovative methods will be needed to provide healthcare practitioners with the ability to interpret genomic data and make evidence-based recommendations. Research is needed to establish appropriate competencies and on making the necessary educational

opportunities available to all healthcare providers effectively, appropriately, and in culturally and linguistically relevant ways across diverse patient populations. Point-of-care clinical decision-support processes are also required. The challenge will be to develop models that can be implemented at the time, place and knowledge level needed to provide effective care.

Equally important is a well-informed public that is supportive of genomics research and appreciates the value of research participation. Consumers will need tools to assess the promises and claims of genomic testing services. Development and implementation of appropriate healthcare policies will depend on educated policy makers. Research is needed to determine the knowledge necessary for making genomically informed clinical decisions at both the individual and societal levels. A variety of pilot efforts should be developed, tested and assessed for their effectiveness in engendering genomically (and more broadly, scientifically and statistically) literate healthcare providers, patients and the general public.

Increasing access to genomic medicine

Genomics will only achieve its full potential to improve health when the advances it engenders become accessible to all. The development of novel and effective mechanisms for involving diverse stakeholder groups is needed to maximize the relevance of genomics to different healthcare systems.

Many existing healthcare infrastructures are poorly suited for the delivery of genomic medicine to all segments of the population. Optimal models for ensuring that the best practices in genomic medicine become available to all at-risk patient populations have yet to be defined. Some possibilities for new approaches include reliance on non-geneticist healthcare providers guided by informatics support, increased use of telemedicine and enhanced genomics education for future generations of healthcare providers. All of these must be pursued.

Concluding comments

The discussions in the 1980s that led to the HGP were motivated by a vision that knowing the human genome sequence would be extraordinarily useful for understanding human biology and disease. For example, Dulbecco wrote⁹⁸ in 1986 that "If we wish to learn more about cancer, we must now concentrate on the cellular genome," and he advocated sequencing "the whole genome of a selected animal species," specifically, the human genome. In 1988, a US National Research Council (NRC) report³ articulated a bold plan for an effort that would culminate in sequencing the human genome; the report stated that such a "project would greatly increase our understanding of human biology and allow rapid progress to occur in the diagnosis and ultimate control of many human diseases." In the past quarter-century, the prescience of this audacious vision has been confirmed. Progress in genomics has been monumental. Although staggering challenges remain, the fundamental goals have not changed—genomics and related large-scale biological studies will, in ways not previously available, lead to a profound understanding about the biology of genomes and disease, to unimaginable advances in medical science, and to powerful new ways for improving human health.

Achieving these goals will continue to rely on new technologies, large-scale collaborative efforts, multidisciplinary and international teams, comprehensiveness, high-throughput data production and analysis, computational intensity, high standards for data quality, rapid data release, and attention to societal implications. The perfusion of genomics into other areas of biomedical research will enable these disciplines to make advances far beyond what is possible today. Achieving such a pervasive positive influence on biomedicine is one of the most gratifying aspects of genomics, as anticipated by the NRC report's detailed 'call to action' blueprint for the HGP³. It is thus with a continuing sense of wonder, a continuing need for urgency, a continuing desire to balance ambition with reality, and a continuing responsibility to protect individuals while maximizing the societal benefits of genomics that we have discussed here some of the many compelling opportunities and significant challenges for the next decade of genomics research. This new vision is ambitious and far-reaching, both in scope and timing. It goes

well beyond what any one organization can realistically support, and will (once again) require the creative energies and expertise of genome scientists around the world and from all sectors, including academic, government and commercial.

Successfully navigating a course from the base pairs of the human genome sequence to the bedside of patients seems within reach, would usher in an era of genomic medicine, would fulfil the promise originally envisioned for the HGP and, most importantly, would benefit all humankind.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. National Research Council. *Mapping and Sequencing the Human Genome* (National Academy Press, 1988).
4. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
5. Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
6. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
7. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Rev. Genet.* **11**, 685–696 (2010).
8. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
9. Pao, W. et al. EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101**, 13306–13311 (2004).
10. Yan, H. et al. *IDH1* and *IDH2* mutations in gliomas. *N. Engl. J. Med.* **360**, 765–773 (2009).
11. Brooke, B. S. et al. Angiotensin II blockade and aortic-root dilation in Marfan’s syndrome. *N. Engl. J. Med.* **358**, 2787–2795 (2008).
12. D’Hulst, C. & Kooy, R. F. Fragile X syndrome: from molecular genetics to therapy. *J. Med. Genet.* **46**, 577–584 (2009).
13. Dietz, H. C. New therapeutic approaches to mendelian disorders. *N. Engl. J. Med.* **363**, 852–863 (2010).
14. Miller, D. T. et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
15. Frueh, F. W. et al. Pharmacogenomic biomarker information in drug labels approved by the United States Food and Drug Administration: prevalence of related drug use. *Pharmacotherapy* **28**, 992–998 (2008).
16. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* doi:10.1038/nature09792 (this issue).
17. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
18. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
19. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
20. Thorisson, G. A. & Stein, L. D. The SNP Consortium website: past, present and future. *Nucleic Acids Res.* **31**, 124–127 (2003).
21. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
22. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
23. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
24. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
25. Gerstein, M. B. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
26. The modENCODE Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
27. Jain, K. K. Applications of nanobiotechnology in clinical diagnostics. *Clin. Chem.* **53**, 2002–2009 (2007).
28. Oita, I. et al. Microfluidics in macro-biomolecules analysis: macro inside in a nano world. *Anal. Bioanal. Chem.* **398**, 239–264 (2010).
29. Metzker, M. L. Sequencing technologies—the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
30. Mardis, E. R. A decade’s perspective on DNA sequencing technology. *Nature* doi:10.1038/nature09796 (this issue).
31. Hiraki, S. & Green, N. S. Newborn screening for treatable genetic conditions: past, present and future. *Obstet. Gynecol. Clin. North Am.* **37**, 11–21 (2010).
32. Levy, P. A. An overview of newborn screening. *J. Dev. Behav. Pediatr.* **31**, 622–631 (2010).
33. Berglund, L. et al. A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* **7**, 2019–2027 (2008).
34. Bjorling, E. & Uhlen, M. Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol. Cell. Proteomics* **7**, 2028–2037 (2008).
35. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nature Rev. Genet.* **11**, 559–571 (2010).
36. Hindorf, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
37. Margulies, E. H. et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
38. Austin, C. P. et al. The Knockout Mouse Project. *Nature Genet.* **36**, 921–924 (2004).
39. Feinberg, A. P. Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nature Biotechnol.* **28**, 1049–1052 (2010).
40. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nature Biotechnol.* **28**, 1057–1068 (2010).
41. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnol.* **28**, 1045–1048 (2010).
42. Rioux, J. D. et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39**, 596–604 (2007).
43. Van Limbergen, J., Wilson, D. C. & Satsangi, J. The genetics of Crohn’s disease. *Annu. Rev. Genomics Hum. Genet.* **10**, 89–116 (2009).
44. Brest, P. et al. Autophagy and Crohn’s disease: at the crossroads of infection, inflammation, immunity, and cancer. *Curr. Mol. Med.* **10**, 486–502 (2010).
45. Cadwell, K. et al. A key role for autophagy and the autophagy gene *Atg16l1* in mouse and human intestinal Paneth cells. *Nature* **456**, 259–263 (2008).
46. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
47. Smushkin, G. & Vella, A. Genetics of type 2 diabetes. *Curr. Opin. Clin. Nutr. Metab. Care* **13**, 471–477 (2010).
48. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
49. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
50. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
51. Couzin-Frankel, J. Ethics. DNA returned to tribe, raising questions about consent. *Science* **328**, 558 (2010).
52. Stover, P. J., Harlan, W. R., Hammond, J. A., Hendershot, T. & Hamilton, C. M. PhenX: a toolkit for interdisciplinary genetics research. *Curr. Opin. Lipidol.* **21**, 136–140 (2010).
53. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
54. Manolio, T. A. & Collins, R. Enhancing the feasibility of large cohort studies. *J. Am. Med. Assoc.* **304**, 2290–2291 (2010).
55. Collins, F. S. The case for a US prospective cohort study of genes and environment. *Nature* **429**, 475–477 (2004).
56. Gulcher, J. R. & Stefansson, K. The Icelandic Healthcare Database and informed consent. *N. Engl. J. Med.* **342**, 1827–1830 (2000).
57. Hudson, K. L., Holohan, M. K. & Collins, F. S. Keeping pace with the times—the Genetic Information Nondiscrimination Act of 2008. *N. Engl. J. Med.* **358**, 2661–2663 (2008).
58. Ludman, E. J. et al. Glad you asked: participants’ opinions of re-consent for dbGap data submission. *J. Empir. Res. Hum. Res. Ethics* **5**, 9–16 (2010).
59. McGuire, A. L., Caulfield, T. & Cho, M. K. Research ethics and the challenge of whole-genome sequencing. *Nature Rev. Genet.* **9**, 152–156 (2008).
60. Wolf, S. M. et al. Managing incidental findings in human subjects research: analysis and recommendations. *J. Law Med. Ethics* **36**, 219–248 (2008).
61. Kohane, I. S. & Taylor, P. L. Multidimensional results reporting to participants in genomic studies: getting it right. *Sci. Transl. Med.* **2**, 37cm19 (2010).
62. Harris, L. et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J. Clin. Oncol.* **25**, 5287–5312 (2007).
63. Allegra, C. J. et al. American Society of Clinical Oncology provisional clinical opinion: testing for *KRAS* gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *J. Clin. Oncol.* **27**, 2091–2096 (2009).
64. Panel on Antiretroviral Guidelines for Adults and Adolescents. *Guidelines for the Use of Antiretroviral Agents in HIV-1 Infected Adults and Adolescents* 1–161 (<http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf>) (Department of Health and Human Services, 2009).
65. Burke, W., Laberge, A. M. & Press, N. Debating clinical utility. *Public Health Genomics* **13**, 215–223 (2010).
66. Shiwa, K. et al. Difficulty in losing weight by behavioral intervention for women with Trp64Arg polymorphism of the β_3 -adrenergic receptor gene. *Int. J. Obes. Relat. Metab. Disord.* **27**, 1028–1036 (2003).
67. Feldstein Ewing, S. W., LaChance, H. A., Bryan, A. & Hutchison, K. E. Do genetic and individual risk factors moderate the efficacy of motivational enhancement therapy? Drinking outcomes with an emerging adult sample. *Addict. Biol.* **14**, 356–365 (2009).
68. McBride, C. M. et al. Future health applications of genomics: priorities for communication, behavioral, and social sciences research. *Am. J. Prev. Med.* **38**, 556–565 (2010).
69. Chan, J. N., Nislow, C. & Emili, A. Recent advances and method development for drug target identification. *Trends Pharmacol. Sci.* **31**, 82–88 (2010).
70. Druker, B. J. et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).

71. Druker, B. J. *et al.* Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med.* **344**, 1038–1042 (2001).
72. Brufsky, A. Trastuzumab-based therapy for patients with HER2-positive breast cancer: from early scientific development to foundation of care. *Am. J. Clin. Oncol.* **33**, 186–195 (2010).
73. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).
74. Sawyers, C. Targeted cancer therapy. *Nature* **432**, 294–297 (2004).
75. Stuart, D. & Sellers, W. R. Linking somatic genetic alterations in cancer to therapeutics. *Curr. Opin. Cell Biol.* **21**, 304–310 (2009).
76. Grasemann, H. & Ratjen, F. Emerging therapies for cystic fibrosis lung disease. *Expert Opin. Emerg. Drugs* **15**, 653–659 (2010).
77. Simon, R. The use of genomics in clinical trial design. *Clin. Cancer Res.* **14**, 5984–5993 (2008).
78. Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947–957 (2009).
79. Caskey, C. T. Using genetic diagnosis to determine individual therapeutic utility. *Annu. Rev. Med.* **61**, 1–15 (2010).
80. Henderson, G. E. *et al.* Clinical trials and medical care: defining the therapeutic misconception. *PLoS Med.* **4**, e324 (2007).
81. Tozzi, V. Pharmacogenetics of antiretrovirals. *Antiviral Res.* **85**, 190–200 (2010).
82. Goetz, M. P. *et al.* The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. *Breast Cancer Res. Treat.* **101**, 113–121 (2007).
83. Angiolillo, D. J. *et al.* Contribution of gene sequence variations of the hepatic cytochrome P450 3A4 enzyme to variability in individual responsiveness to clopidogrel. *Arterioscler. Thromb. Vasc. Biol.* **26**, 1895–1900 (2006).
84. Rieder, M. J. *et al.* Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352**, 2285–2293 (2005).
85. Sankar, P. *et al.* Genetic research and health disparities. *J. Am. Med. Assoc.* **291**, 2985–2989 (2004).
86. Ramos, E. & Rotimi, C. The A's, G's, C's, and T's of health disparities. *BMC Med. Genomics* **2**, 29 (2009).
87. Rotimi, C. N. & Jorde, L. B. Ancestry and disease in the age of genomic medicine. *N. Engl. J. Med.* **363**, 1551–1558 (2010).
88. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA* **103**, 14068–14073 (2006).
89. Kao, W. H. *et al.* MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature Genet.* **40**, 1185–1192 (2008).
90. Race, Ethnicity, and Genetics Working Group. The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.* **77**, 519–532 (2005).
91. Murphy, S. *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* **19**, 1675–1681 (2009).
92. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
93. Del Fiol, G. *et al.* Integrating genetic information resources with an EHR. *AMIA Annu. Symp. Proc.* **2006**, 904 (2006).
94. Buntin, M. B., Jain, S. H. & Blumenthal, D. Health information technology: laying the infrastructure for national health reform. *Health Aff. (Millwood)* **29**, 1214–1219 (2010).
95. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
96. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
97. Loukides, G., Gkoulalas-Divanis, A. & Malin, B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl Acad. Sci. USA* **107**, 7898–7903 (2010).
98. Dulbecco, R. A turning point in cancer research: sequencing the human genome. *Science* **231**, 1055–1056 (1986).
99. Toronto International Data Release Workshop Authors. Prepublication data sharing. *Nature* **461**, 168–170 (2009).
100. Contreras, J. L. Information access. Prepublication data release, latency, and genome commons. *Science* **329**, 393–394 (2010).
101. Robinson, C. J., Bohannan, B. J. & Young, V. B. From structure to function: the ecology of host-associated microbial communities. *Microbiol. Mol. Biol. Rev.* **74**, 453–476 (2010).
102. The NIH HMP Working Group. The NIH Human Microbiome Project. *Genome Res.* **19**, 2317–2323 (2009).
103. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
104. Kuczynski, J. *et al.* Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.* **11**, 210 (2010).
105. Stein, L. D. The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207 (2010).

Acknowledgements This paper is endorsed by the National Advisory Council for Human Genome Research, whose members (including ad hoc members (indicated by an asterisk) at the September 2010 meeting) are M. Boehnke, M. Chee, R. Chisholm, J. Contreras, R. Cooper, C. Fraser-Liggett, R. Gibbs, G. Ginsberg, R. Hardison*, H. McLeod*, J. Mesirov*, R. Myers, P. O'Rourke, P. Ossorio, P. Sankar*, D. Valle, R. Weinshilboum, D. Williams* and R. Wilson*. The dedicated effort of NHGRI staff was instrumental for the strategic planning process that culminated in this paper. The final stages of this process were led by E. Green and M. Guyer, and involved extensive contributions from K. Wetterstrand, A. Bailey and S. Vasquez. We also thank the additional NHGRI staff who made important contributions to the writing of this paper: C. Austin, L. Biesecker, V. Bonazzi, V. Bonham, J. Boyer, L. Brody, L. Brooks, G. Feero, E. Feingold, A. Felsenfeld, P. Frosst, P. Good, B. Graham, D. Hadley, S. Hart, L. Hindorff, C. McBride, J. McEwen, M. Muenke, J. Mullikin, E. Ostrander, A. Pillai, R. Pozzatti, L. Proctor, E. Ramos, L. Rodriguez, J. Schloss, J. Segre, J. Struewing, L. Thompson and A. Wilson. We thank the following for editorial input: D. Altshuler, M. Boehnke, D. Botstein, J. Evans, C. Fraser-Liggett, E. Lander, R. Lifton, R. Myers, D. Valle, R. Waterston and H. Varmus. We acknowledge the graphical contributions of D. Leja and J. Latman, as well as the logistical support of P. Klein, J. Mullaney and E. Rolfes. We also thank H. Dietz and S. Terry who, along with some of those previously listed, crafted topical white papers for the strategic planning process (see <http://genome.gov/Planning>). Finally, we thank the many participants at various NHGRI strategy meetings, as well as F. Collins and A. Guttman for their early leadership of this strategic planning process.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to E.D.G. (egreen@nhgri.nih.gov).

National Human Genome Research Institute

Overall leadership Eric D. Green¹ & Mark S. Guyer¹

Coordination of writing contributions (see Acknowledgements for list of other contributors) Teri A. Manolio¹ & Jane L. Peterson¹

¹National Human Genome Research Institute, National Institutes of Health, 31 Center Dr., Bethesda, Maryland 20892-2152, USA.

The genomic complexity of primary human prostate cancer

Michael F. Berger^{1†*}, Michael S. Lawrence^{1*}, Francesca Demichelis^{2,3*}, Yotam Drier^{4*}, Kristian Cibulskis¹, Andrey Y. Sivachenko¹, Andrea Sboner^{5,6}, Raquel Esgueva², Dorothee Pflueger², Carrie Sougnez¹, Robert Onofrio¹, Scott L. Carter¹, Kyung Park², Lukas Habegger⁶, Lauren Ambrogio¹, Timothy Fennell¹, Melissa Parkin¹, Gordon Saksena¹, Douglas Voet¹, Alex H. Ramos^{1,7}, Trevor J. Pugh^{1,7,8}, Jane Wilkinson¹, Sheila Fisher¹, Wendy Winckler¹, Scott Mahan¹, Kristin Ardlie¹, Jennifer Baldwin¹, Jonathan W. Simons⁹, Naoki Kitabayashi², Theresa Y. MacDonald², Philip W. Kantoff^{7,8}, Lynda Chin^{1,7,8,10}, Stacey B. Gabriel¹, Mark B. Gerstein^{5,6,11}, Todd R. Golub^{1,12,13,14}, Matthew Meyerson^{1,7,8,14}, Ashutosh Tewari¹⁵, Eric S. Lander^{1,7,16}, Gad Getz¹, Mark A. Rubin² & Levi A. Garraway^{1,7,8,14}

Prostate cancer is the second most common cause of male cancer deaths in the United States. However, the full range of prostate cancer genomic alterations is incompletely characterized. Here we present the complete sequence of seven primary human prostate cancers and their paired normal counterparts. Several tumours contained complex chains of balanced (that is, 'copy-neutral') rearrangements that occurred within or adjacent to known cancer genes. Rearrangement breakpoints were enriched near open chromatin, androgen receptor and ERG DNA binding sites in the setting of the ETS gene fusion *TMPRSS2-ERG*, but inversely correlated with these regions in tumours lacking ETS fusions. This observation suggests a link between chromatin or transcriptional regulation and the genesis of genomic aberrations. Three tumours contained rearrangements that disrupted *CADM2*, and four harboured events disrupting either *PTEN* (unbalanced events), a prostate tumour suppressor, or *MAGI2* (balanced events), a *PTEN* interacting protein not previously implicated in prostate tumorigenesis. Thus, genomic rearrangements may arise from transcriptional or chromatin aberrancies and engage prostate tumorigenic mechanisms.

Among men in the United States, prostate cancer accounts for more than 200,000 new cancer cases and 32,000 deaths annually¹. Although androgen deprivation therapy yields transient efficacy, most patients with metastatic prostate cancer eventually die of their disease. These aspects underscore the critical need to articulate both genetic underpinnings and novel therapeutic targets in prostate cancer.

Recent years have heralded a marked expansion in our understanding of the somatic genetic basis of prostate cancer. Of considerable importance has been the discovery of recurrent gene fusions that render ETS transcription factors under the control of androgen-responsive or other promoters^{2–5}. These findings suggest that genomic rearrangements may comprise a major mechanism driving prostate carcinogenesis. Other types of somatic alterations also engage important mechanisms^{6–8}; however, the full spectrum of prostate cancer genomic alterations remains incompletely characterized. Moreover, although the androgen signalling axis represents an important therapeutic focal point^{9,10}, relatively few additional drug targets have yet been elaborated by genetic studies of prostate cancer¹¹. To discover additional genomic alterations that may underpin lethal prostate cancer, we performed paired-end, massively parallel sequencing on tumour and matched normal genomic DNA obtained from seven patients with 'high-risk' primary prostate cancer.

Landscape of genomic alterations

All patients harboured tumours of stage T2c or greater, and Gleason grade 7 or higher. Serum prostate-specific antigen levels ranged from 2.1 to 10.2 ng ml⁻¹ (Supplementary Table 1). Three tumours contained chromosomal rearrangements involving the *TMPRSS2* (transmembrane protease, serine 2)–*ERG* (v-ets erythroblastosis virus E26 oncogene homologue (avian)) loci as determined by fluorescence *in situ* hybridization (FISH) and PCR with reverse transcription (RT-PCR)² (Table 1 and Supplementary Table 1). We obtained approximately 30-fold mean sequence coverage for each sample, and reliably detected somatic mutations in more than 80% of the genome (described in Supplementary Information). Circos plots¹² indicating genomic rearrangements and copy number alterations for each prostate cancer genome are shown in Fig. 1.

We identified a median of 3,866 putative somatic base mutations (range 3,192–5,865) per tumour (Supplementary Table 2); the estimated mean mutation frequency was 0.9 per megabase (see Supplementary Methods). This mutation rate is similar to that observed in acute myeloid leukaemia and breast cancer^{13–16} but 7–15-fold lower than rates reported for small cell lung cancer and melanoma^{17–19}. The mutation rate at CpG (that is, cytosine–phosphate–guanine) dinucleotides was more than tenfold higher than at all other genomic positions

¹The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. ²Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, New York 10065, USA. ³Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York 10021, USA. ⁴Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. ⁵Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. ⁶Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. ⁷Harvard Medical School, Boston, Massachusetts 02115, USA. ⁸Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ⁹The Prostate Cancer Foundation, Santa Monica, California 90401, USA. ¹⁰Belfer Institute for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹¹Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. ¹²Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹³Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. ¹⁴Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹⁵Department of Urology, Institute of Prostate Cancer and Lefrak Center of Robotic Surgery, Weill Cornell Medical College and New York Presbyterian Hospitals, New York, New York 10065, USA. ¹⁶Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [†]Present address: Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA.

*These authors contributed equally to this work.

Table 1 | Landscape of somatic alterations in primary human prostate cancers

	Tumour						
	PR-0508	PR-0581*	PR-1701*	PR-1783	PR-2832*	PR-3027	PR-3043
Tumour bases sequenced	97.8×10^9	93.9×10^9	110×10^9	90.9×10^9	106×10^9	93.6×10^9	94.9×10^9
Normal bases sequenced	96.7×10^9	57.8×10^9	108×10^9	92.3×10^9	103×10^9	87.8×10^9	96.6×10^9
Tumour haploid coverage	31.8	30.5	35.8	29.5	34.4	30.4	30.8
Normal haploid coverage	31.4	18.8	34.9	30.0	33.4	28.5	31.4
Callable fraction	0.84	0.83	0.87	0.82	0.84	0.84	0.85
Estimated tumour purity†	0.73	0.60	0.49	0.75	0.59	0.74	0.68
All point mutations (high confidence)	3,898 (1,447)	3,829 (1,430)	3,866 (1,936)	4,503 (2,227)	3,465 (1,831)	5,865 (2,452)	3,192 (1,713)
Non-silent coding mutations (high confidence)	16 (5)	20 (3)	24 (9)	32 (20)	13 (7)	43 (16)	14 (10)
Mutation rate per Mb	0.7	0.7	0.8	1.0	0.8	1.2	0.7
Rearrangements	53	67	90	213	133	156	43

* Harbours *TMPS2-ERG* gene fusion

† Estimated from SNP array-derived allele specific copy number levels using the ABSOLUTE algorithm (Supplementary Methods).

(Supplementary Fig. 1). A median of 20 non-synonymous base mutations per sample were called within protein-coding genes (range 13–43; Supplementary Table 3). We also identified six high-confidence coding indels (4 deletions, 2 insertions) ranging from 1 to 9 base pairs (bp) in length, including a 2-bp frameshift insertion in the tumour suppressor gene, *PTEN* (phosphatase and tensin homologue; Supplementary Table 4, Supplementary Fig. 2).

Two genes (*SPTA1* and *SPOP*) harboured mutations in two out of seven tumours. *SPTA1* encodes a scaffold protein involved in erythroid cell shape specification, while *SPOP* encodes a modulator of Daxx-mediated ubiquitination and transcriptional regulation²⁰. The *SPOP* mutations exceeded the expected background rate in these tumours ($Q = 0.055$). (Q is defined as the false discovery rate (FDR)-corrected P value.) Moreover, *SPOP* was also found significantly mutated in a separate study of prostate cancer²¹. Interestingly, the chromatin modifiers *CHD1*, *CHD5* and *HDAC9* were mutated in 3 out of 7 prostate cancers. These genes regulate embryonic stem cell pluripotency, gene regulation, and tumour suppression^{22–24}. Members of the HSP-1 stress response complex (*HSPA2*, *HSPA5* and *HSP90AB1*) were also mutated in three out of seven tumours. The corresponding proteins form a chaperone complex targeted by several anticancer drugs in development²⁵. Furthermore, we found a single KEGG pathway ‘antigen processing and presentation’ to be significantly mutated out of 616 diverse gene sets corresponding to gene families and known pathways ($Q = 0.0021$). This result is intriguing, given the clinical benefit associated with immunotherapy for prostate cancer^{26,27}. Other known

cancer genes were mutated in single tumours, including *PRKCI* and *DICER*. Thus, some coding mutations may contribute to prostate tumorigenesis and suggest possible therapeutic interventions.

Complex patterns of balanced rearrangements

Given the importance of oncogenic gene fusions in prostate cancer, we next characterized the spectrum of chromosomal rearrangements. We identified a median of 90 rearrangements per genome (range 43–213) supported by ≥ 3 distinct read pairs (Supplementary Table 5). This distribution of rearrangements was similar to that previously described for breast cancer²⁸. We examined 594 candidate rearrangements by multiplexed PCR followed by massively parallel sequencing, and validated 78% of events by this approach (Supplementary Methods). Three genes disrupted by rearrangements also harboured non-synonymous mutations in another sample: *ZNF407*, *CHD1* and *PTEN*. Notably, the chromatin modifier *CHD1*, which contains a validated splice site mutation in prostate tumour PR-1701 (as indicated above), also harboured intragenic breakpoints in two additional samples (PR-0508 and PR-1783). These rearrangements predict truncated proteins, raising the possibility that dysregulated *CHD1* may contribute to a block in differentiation in some prostate cancer precursor cells²².

In 88% of cases, the fusion point could be mapped to base pair resolution (Supplementary Methods). The most common type of fusion involved a precise join, with neither overlapping nor intervening sequence at the rearrangement junction. In a minority of cases, an overlap (microhomology) of 2 bp or more was observed. The

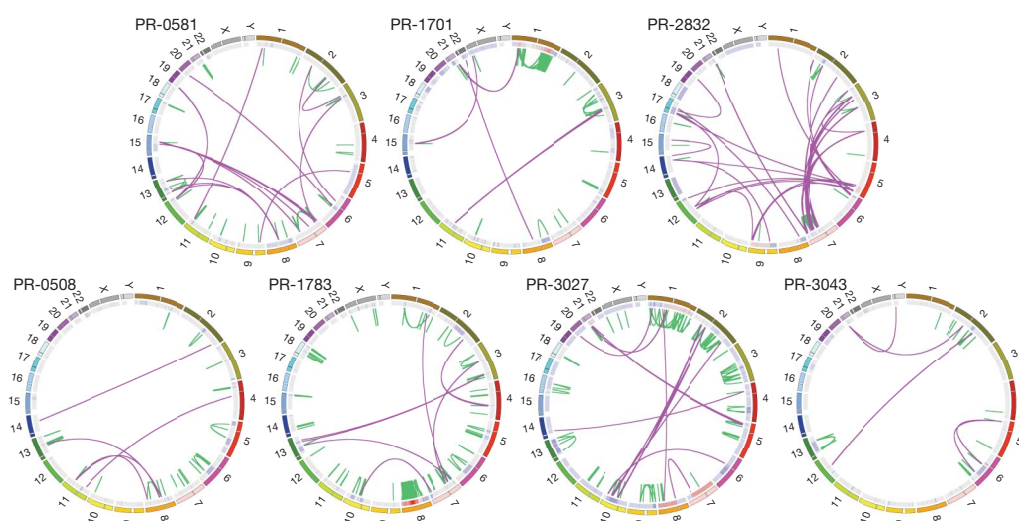


Figure 1 | Graphical representation of seven prostate cancer genomes. Each Circos plot¹² depicts the genomic location in the outer ring and chromosomal copy number in the inner ring (red, copy gain; blue, copy loss). Interchromosomal translocations and intrachromosomal rearrangements are

shown in purple and green, respectively. Genomes are organized according to the presence (top row) or absence (bottom row) of the *TMPS2-ERG* gene fusion.

rearrangement frequency declined by approximately twofold for each base of microhomology. This result differed from the patterns seen in breast tumours, in which the most common junction involved a microhomology of 2–3 bp (ref. 28). Thus, mechanisms by which rearrangements are generated may differ between prostate and breast cancer.

Detailed examination of these chromosomal rearrangements revealed a distinctive pattern of balanced breaking and rejoining not previously observed in solid tumours: several genomes contained complex inter- and intra-chromosomal events involving an exchange of 'breakpoint arms'. A mix of chimaeric chromosomes was thereby generated, without concomitant loss of genetic material (that is, all breakpoints produced balanced translocations, illustrated conceptually in Fig. 2a).

This 'closed chain' pattern of breakage and rejoining was evident in each of the *TMPRSS2-ERG* fusion-positive prostate cancers. In two such cases, both the *TMPRSS2* and *ERG* genomic loci were involved in a closed chain of breakpoints. For example, the *TMPRSS2-ERG* gene

fusion in PR-1701 was produced by a closed quartet of balanced translocations on chromosomes 21 and 1 (Fig. 2b). The *TMPRSS2-ERG* gene fusion in PR-0581 occurred within a closed trio of intra-chromosomal rearrangements involving *C21ORF45*, *ERG* and *TMPRSS2* (Supplementary Fig. 3).

One noteworthy closed chain of rearrangements harboured breakpoints situated independently of *TMPRSS2-ERG* (Supplementary Fig. 4) but in close proximity to multiple known cancer genes or orthologues. This chain (found in sample PR-2832) contained breakpoint pairs at the following loci: (1) 60 bp from exon 6 of TANK binding kinase 1 (*TBK1* or 'NF- κ B-activating kinase')²⁹; (2) within the first intron of *TP53* (7 kilobases (kb) upstream of translation start); (3) 51 kb from *MAP2K4* (a kinase recently shown to induce anchorage-independent growth via mutations²¹); and (4) 3 kb from the *ABL1* proto-oncogene (Fig. 2c). This striking phenomenon suggests that complex translocations may dysregulate multiple genes in parallel to drive prostate tumorigenesis.

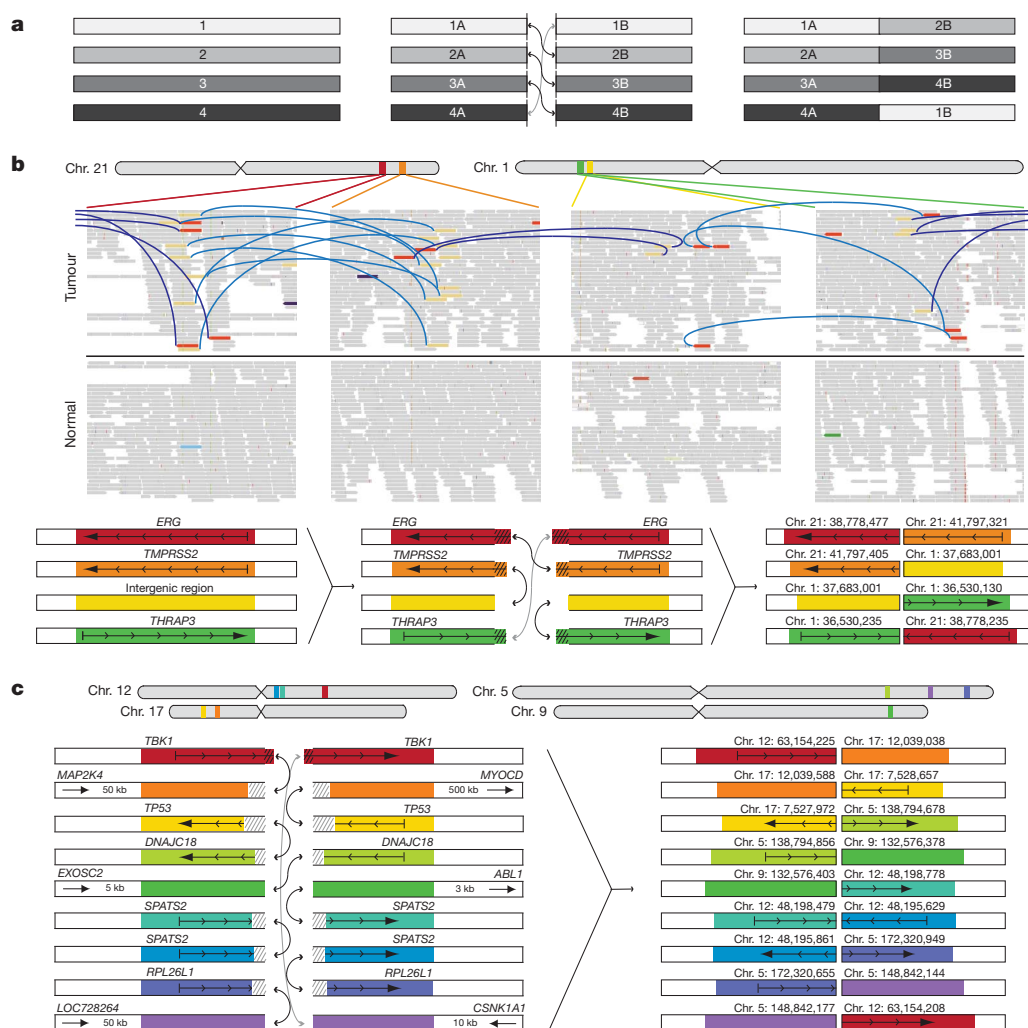


Figure 2 | Complex structural rearrangements in prostate cancer.

a, Diagram of 'closed chain' pattern of chromosomal breakage and rejoining. Breaks are induced in a set of loci (left), followed by an exchange of free ends without loss of chromosomal material (middle), leading to the observed pattern of balanced (copy neutral) translocations involving a closed set of breakpoints (right). **b**, Complex rearrangement in prostate tumour PR-1701. *TMPRSS2-ERG* is produced by a closed quartet of balanced rearrangements involving 4 loci on chromosomes 1 and 21. Top, each rearrangement is supported by the presence of discordant read pairs in the tumour genome but not the normal genome (coloured bars connected by blue lines). Thin bars represent sequence reads; directionality represents mapping orientation on the reference genome.

Figures are based on the Integrative Genomics Viewer (<http://www.broadinstitute.org/igv>). Bottom, Diagram of breakpoints and balanced translocations. Hatching indicates sequences that are duplicated in the derived chromosomes at the resulting fusion junctions. **c**, Complex rearrangement in prostate tumour PR-2832 involving breakpoints and fusions at 9 distinct genomic loci. Hatching indicates sequences that are duplicated or deleted in the derived chromosomes at the resulting fusion junctions. For breakpoints in intergenic regions, the nearest gene in each direction is shown. In addition to the sheer number of regions involved, this complex rearrangement is notable for the abundance of breakpoints in or near cancer-related genes, such as *TBK1*, *MAP2K4*, *TP53* and *ABL1*.

Association of rearrangements and epigenetic marks

The closed chain pattern of chromosomal breakpoints also raised the possibility that multiple genomic regions might become spatially co-localized before undergoing rearrangement. Conceivably, such a phenomenon could reflect migration to 'transcription factories'—preassembled nuclear subcompartments that contain RNA polymerase II holoenzyme³⁰. In prostate cells, androgen signalling has been shown to induce co-localization of *TMPRSS2* and *ERG*, thereby allowing double-strand breaks to facilitate gene fusion formation^{31–33}. A role for transcription in the genesis of *TMPRSS2-ERG* in PR-1701 seems plausible, as genomic sequences of up to 240 bp are duplicated at the resulting fusion junctions (Fig. 2b). Alternatively, chains of breakpoints might reflect the clustering of active and inactive chromatin within the recently demonstrated fractal globule structure of nuclear architecture³⁴. Stimulated by these models, we considered whether the genomic regions involved in prostate cancer rearrangements exhibited similarities in terms of either transcriptional patterns or chromatin marks. Here, we used published chromatin immunoprecipitation and massively parallel sequencing (ChIP-seq) data from VCaP, an androgen-sensitive prostate cancer cell line that harbours the *TMPRSS2-ERG* gene fusion³⁵.

The location of rearrangement breakpoints from the *TMPRSS2-ERG* fusion-positive tumour PR-2832 showed significant spatial correlation with various marks of open chromatin in VCaP cells (Fig. 3 and Supplementary Fig. 5). These marks included ChIP-seq peaks corresponding to RNA polymerase II (pol II, $P = 1.0 \times 10^{-15}$), histone H3K4 trimethylation (H3K4me3, $P = 3.1 \times 10^{-7}$), histone H3K36 trimethylation (H3K36me3, $P = 3.5 \times 10^{-12}$) and histone H3 acetylation (H3ace, $P = 9.5 \times 10^{-12}$) (Fig. 3). Similar statistical correlations were observed for peaks corresponding to the androgen receptor (AR) ($P = 1.1 \times 10^{-5}$) and ERG binding sites ($P = 4.9 \times 10^{-14}$) (Fig. 3 and

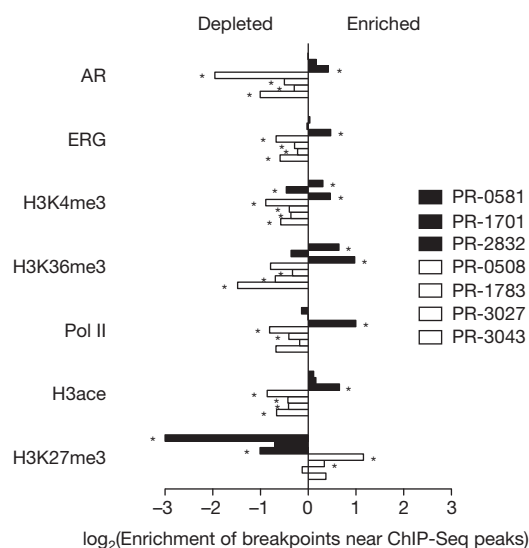


Figure 3 | Association between rearrangement breakpoints and genome-wide transcriptional/histone marks in prostate cancer. ChIP-seq binding peaks were defined previously for the *TMPRSS2-ERG* positive (ERG-positive) prostate cancer cell line VCaP³⁵. For each genome, enrichment of breakpoints within 50 kb of each set of binding peaks was determined relative to a coverage-matched simulated background (see Methods). *TMPRSS2-ERG*-positive prostate tumours are in black; ETS fusion-negative prostate tumours are in white. Enrichment is displayed as the ratio of the observed breakpoint rate to the background rate near each indicated set of ChIP-seq peaks. Rearrangements in ETS fusion-negative tumours are depleted near marks of active transcription (AR, ERG, H3K4me3, H3K36me3, Pol II and H3ace) and enriched near marks of closed chromatin (H3K27me3). *P*-values were calculated according to the binomial distribution and are displayed in Supplementary Fig. 5 and Supplementary Table 6. *Significant associations passing a false discovery rate cut-off of 5%.

Supplementary Table 6), consistent with the substantial overlap between AR and ERG binding locations in VCaP cells³⁵. (We did not observe significant enrichment of either AR or ERG binding site sequences in the vicinity of these breakpoints.) In the other ERG fusion-positive tumours (PR-0581 or PR-1701), the correlations between breakpoints and ChIP-seq peaks were intermittently apparent, albeit much less significant.

Rearrangement breakpoints from all four ETS fusion-negative tumours were inversely correlated with these same marks of open chromatin and AR/ERG binding (Fig. 3 and Supplementary Fig. 5). In fact, breakpoints from two of four ETS-negative tumours were significantly correlated with marks of histone H3K27 trimethylation (H3K27me3) in VCaP cells, which denote inactive chromatin and transcriptional repression (Fig. 3). This result suggested that somatic rearrangements might occur within closed chromatin in some tumour cells, or that the epigenetic architecture or transcriptional program of some *TMPRSS2-ERG* fusion-positive cells differs markedly from that of ERG fusion-negative cells. In support of the former, we observed a similar enrichment of PR-2832 rearrangements and depletion of fusion-negative rearrangements near marks of active transcription profiled in several additional cell lines, including fusion-negative prostate cancer cell lines LNCaP and PC-3 as well as three cell lines derived from non-prostate lineages (Supplementary Fig. 5)^{35–37}.

On the basis of these results, we performed similar analyses comparing the chromatin state in VCaP cells to rearrangement patterns of other cancer types. No statistically significant correlations or inverse correlations were observed between VCaP ChIP-seq data and rearrangement breakpoints obtained from a melanoma cell line¹⁸, a small-cell lung cancer cell line¹⁷, or a primary non-small-cell lung tumour³⁸ (Supplementary Fig. 5 and Supplementary Table 6). However, rearrangements from 16 out of 18 breast tumours and cell lines examined²⁸ exhibited a pattern of association similar to that observed in prostate tumour PR-2832 (Supplementary Fig. 6). Notably, breakpoints in these tumours were also strongly associated with oestrogen receptor (ER) binding sites derived from the breast cancer cell line MCF-7 (ref. 39). Furthermore, we observed a strong association between ER ChIP-seq peaks from MCF-7 and all VCaP ChIP-seq peaks corresponding to open chromatin, AR and ERG binding ($P < 10^{-90}$; Supplementary Fig. 6). Thus, patterns of open chromatin may be highly overlapping in some hormone-driven cancer cells. Such regions may correlate significantly with sites of somatic rearrangement in cancers of the prostate, breast, and possibly other tissues.

To examine whether processes linked to chromatin reorganization and DNA rearrangement are also associated with increased mutation frequency, we tested for enrichment of point mutations near regions of ChIP-seq peaks and rearrangement breakpoints. We observed a significantly reduced prevalence of point mutations near marks of VCaP active transcription—and slight enrichment of mutations in closed chromatin—in all seven prostate tumours (Supplementary Fig. 7). This pattern is consistent with both negative selection and transcription-coupled DNA repair. Additionally, we observed a significant enrichment of mutations near rearrangement breakpoints in five out of seven prostate tumours (Supplementary Fig. 7). Although the increased rate of mutations near rearrangements may conceivably reflect activation-induced cytosine deaminase in the double strand break repair process^{31,40}, we did not observe a significant overrepresentation of any one class of mutation among those located near breakpoints.

Recurrent rearrangements involving *CADM2*

Sixteen genes harboured a somatic rearrangement in at least two prostate tumours (Supplementary Table 7), and four contained rearrangements in three out of seven tumours. In addition to *TMPRSS2* and *ERG*, the latter included *CSMD3* and *CADM2* (cell adhesion molecule 2). These genes were rearranged at a frequency beyond that expected by chance, even after correcting for gene size (Supplementary Table 8).

CSMD3 encodes a giant gene that contains multiple CUB and sushi repeats. However, we did not observe additional *CSMD3* rearrangements by FISH in an independent analysis of 94 prostate tumours (Supplementary Fig. 8).

CADM2 encodes a nectin-like member of the immunoglobulin-like cell adhesion molecules. Several nectin-like proteins exhibit tumour suppressor properties in various contexts. Analysis of single nucleotide polymorphism (SNP) array-derived copy number profiles of tumours and cell lines^{41,42} suggests that *CADM2* does not reside near a fragile site (Supplementary Fig. 9). At the same time, the complexity of *CADM2* rearrangements (Fig. 4a) suggested that a simple FISH validation approach might prove insufficient to determine the overall frequency of *CADM2* disruption. Nevertheless, we screened an independent cohort of 90 additional prostate tumours using a 'break-apart' FISH assay designed to query the *CADM2* locus (Supplementary Fig. 8). *CADM2* aberrations were detected in 6 out of 90 samples (5 rearrangements and 1 copy gain; Fig. 4b). These results confirmed that *CADM2* is recurrently disrupted in prostate cancer, and they are likely to represent a lower bound for the true prevalence of *CADM2* alteration in this malignancy.

Rearrangements disrupting *PTEN* and *MAGI2*

Two prostate tumours contained breakpoints within the *PTEN* tumour suppressor gene⁶ (Fig. 4c). In both cases, the rearrangements generated heterozygous deletions that were confirmed by FISH analysis (Supplementary Fig. 10). In one tumour (PR-0581), *PTEN* rearrangement co-occurred with a dinucleotide insertion within the *PTEN* coding sequence (described above).

Two additional tumours harboured rearrangements disrupting the *MAGI2* (membrane associated guanylate kinase, WW and PDZ domain containing 2) gene, which encodes a *PTEN*-interacting protein^{43,44} (Fig. 4c). In one tumour (PR-0508), two independent but closely aligned inversion events (marking both ends of a 450-kb inverted sequence) affected the *MAGI2* locus. In the other tumour (PR-2832), two long-range intrachromosomal inversions were observed, raising the possibility of heterogeneous subclones harbouring independent

MAGI2 rearrangements. Thus, four out of seven tumours harboured rearrangements predicted to inactivate *PTEN* or *MAGI2*, including all three tumours harbouring *TMPRSS2-ERG* rearrangements. Although a tumour suppressor function for *MAGI2* has not been established previously, this gene was recently shown to undergo rearrangement in the genome of a melanoma cell line¹⁸, another tumour type in which *PTEN* loss is prevalent. In principle, genomic rearrangements that subvert *PTEN* function either directly or indirectly (for example, through loss of *MAGI2*) might dysregulate the PI3 kinase pathway in prostate cancer.

Whereas both *PTEN* rearrangements involved chromosomal copy loss, the *MAGI2* rearrangements were balanced events (Supplementary Fig. 11). Like *CSMD3* and *CADM2*, *MAGI2* does not appear to reside near a fragile site (Supplementary Fig. 9). We screened 88 independent prostate tumours using FISH inversion probes and identified 3 additional samples harbouring similar inversions, each of which was wild type for *PTEN* disruption (Fig. 4d and Supplementary Fig. 8). As with *CADM2* above, these FISH findings may underestimate the true frequency of *MAGI2* disruption in prostate cancer.

We further analysed the *PTEN* and *MAGI2* loci using high-density SNP arrays obtained from 66 primary prostate cancers. As shown in Supplementary Fig. 11, focal somatic deletions affecting the *PTEN* locus were commonly observed in these tumours, as expected. Interestingly, no somatic copy number alterations were observed at the *MAGI2* locus in either prostate tumour found to contain *MAGI2* rearrangements by genome sequencing (Supplementary Fig. 11). Conceivably, this region may also harbour genes whose loss would be deleterious to prostate cancer cells. More generally, these findings suggest that extensive shotgun paired-end sequencing (as opposed to lower-resolution approaches) may be required to elaborate the compendium of genes targeted by somatic alterations in prostate cancer.

Discussion

This study represents the first whole genome sequencing analysis of human prostate cancer. Systematic genome characterization efforts have often focused primarily on gene-coding regions to identify 'driver'

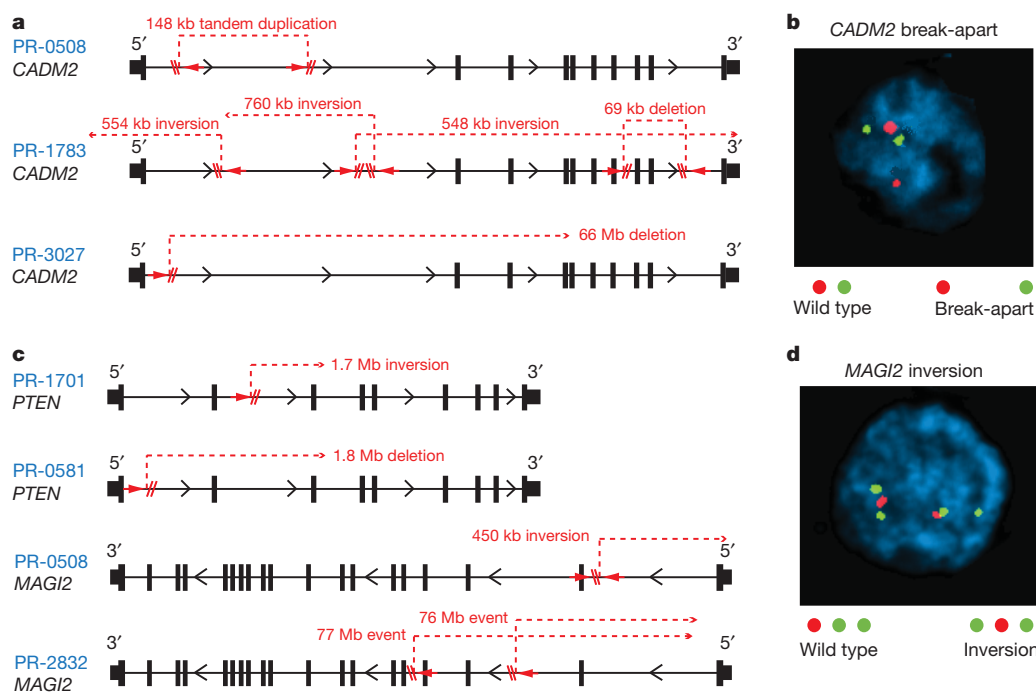


Figure 4 | Disruption of *CADM2* and the *PTEN* pathway by rearrangements. **a**, Location of intragenic breakpoints in *CADM2*. **b**, *CADM2* break-apart demonstrated by FISH in an independent prostate tumour. **c**, Location of intragenic breakpoints in *PTEN* (top) and *MAGI2* (bottom).

d, *MAGI2* inversion demonstrated by FISH in an independent prostate tumour, using probes flanking *MAGI2* (red and green) and an external reference probe also on chromosome 7q (green). The probes and strategy for detecting novel rearrangements by FISH are shown in diagram form in Supplementary Fig. 8.

or 'druggable' alterations^{45–47}. In contrast, the high prevalence of recurrent gene fusions has highlighted chromosomal rearrangements as critical initiating events in prostate cancer^{2,3}. Genome sequencing data indicate that complex rearrangements may enact pivotal gain- and loss-of-function driver events in primary prostate carcinogenesis. Moreover, many rearrangements may occur preferentially in genes that are spatially localized together with transcriptional or chromatin compartments, perhaps initiated by DNA strand breaks and erroneous repair. The complexity of 'closed chain' and other rearrangements suggests that complete genome sequencing—as opposed to approaches focused on exons or gene fusions—may be required to elaborate the spectrum of mechanisms directing prostate cancer genesis and progression.

A positive correlation exists between the location of breakpoints in *TMPRSS2-ERG*-positive tumour cells and open chromatin in VCaP cells, and also between breakpoints present in ETS fusion-negative cells and VCaP regions of closed chromatin. This suggests that breakpoints may preferentially occur within regions of open chromatin in some *TMPRSS2-ERG*-positive tumour cells while raising alternative possibilities for the genesis of breakpoints in ETS fusion-negative cells. Conceivably, somatic rearrangements may occur within regions of closed chromatin in tumour cells lacking ETS gene fusions. Alternately, such tumour cells may have distinct transcriptional or chromatin patterns, with many regions that are closed in VCaP being open in these cells. Clustering of breakpoints within active regions might also reflect selection for functionally consequential rearrangements during tumorigenesis. The relative contribution of these aspects to tumorigenesis will probably be informed by additional integrative analyses of epigenetic and structural genomic data sets across many tumour types.

Previous studies of genetically engineered mouse models have shown that the combination of ERG dysregulation and PTEN loss triggers the formation of aggressive prostate tumours^{48,49}. This same combination identifies a subtype of human prostate cancer characterized by poor prognosis⁵⁰. The discovery of *MAGI2* genomic rearrangements in prostate cancer suggests that interrogating both the *PTEN* and *MAGI2* loci might improve prognostication and patient stratification for clinical trials of PI3 kinase pathway inhibitors. Additional mutated genes discovered in this study also suggest interesting therapeutic avenues. For example, the presence of point mutations involving chromatin modifying genes and the Hsp-1 stress response complex (which includes the Hsp90 chaperone protein targeted by several drugs in development) raises the possibility that these cellular processes may represent targetable dependencies in some prostate tumours. Overall, complete genome sequencing of large numbers of relapsing primary and metastatic prostate cancers promises to define a genetic cartography that assists in tumour classification, elaborates mechanisms of carcinogenesis and identifies new targets for therapeutic intervention.

METHODS SUMMARY

The complete genomes of seven prostate tumours and patient-matched normal samples were sequenced to approximately 30-fold haploid coverage on an Illumina GA II sequencer. DNA was extracted from patient blood and from tumours following radical prostatectomy, and was subjected to extensive quality control procedures to monitor DNA structural integrity, genotype concordance, and tumour purity and ploidy. Standard paired-end libraries (~400-bp inserts) were sequenced as 101-bp paired-end reads. Raw sequencing data were processed by Illumina software and passed to the Picard pipeline, which produced a single BAM file for each sample storing all reads with well-calibrated quality scores together with their alignments to the reference genome. BAM files for each tumour/normal sample pair were analysed by the Firehose pipeline to characterize the full spectrum of somatic mutations in each tumour, including base pair substitutions, short insertions and deletions, and large-scale structural rearrangements. A subset of base pair mutations and rearrangements were validated using independent technologies in order to assess the specificity of the detection algorithms. FISH was also performed for selected recurrent rearrangements. The locations of all rearrangement breakpoints were compared to previously published ChIP binding peaks from related cell types to test for global associations

between rearrangements and a range of epigenetic marks. A complete description of the materials and methods is provided in Supplementary Information.

Received 30 June; accepted 8 December 2010.

- Jemal, A., Siegel, R., Xu, J. & Ward, E. Cancer statistics, 2010. *CA Cancer J. Clin.* **60**, 277–300 (2010).
- Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
- Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595–599 (2007).
- Helgeson, B. E. *et al.* Characterization of *TMPRSS2:ETV5* and *SLC45A3:ETV5* gene fusions in prostate cancer. *Cancer Res.* **68**, 73–80 (2008).
- Tomlins, S. A. *et al.* *TMPRSS2:ETV4* gene fusions define a third molecular subtype of prostate cancer. *Cancer Res.* **66**, 3396–3400 (2006).
- Li, J. *et al.* *PTEN*, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
- Visakorpi, T. *et al.* *In vivo* amplification of the androgen receptor gene and progression of human prostate cancer. *Nature Genet.* **9**, 401–406 (1995).
- Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
- Tran, C. *et al.* Development of a second-generation antiandrogen for treatment of advanced prostate cancer. *Science* **324**, 787–790 (2009).
- Attard, G. *et al.* Phase I clinical trial of a selective inhibitor of CYP17, abiraterone acetate, confirms that castration-resistant prostate cancer commonly remains hormone driven. *J. Clin. Oncol.* **26**, 4563–4571 (2008).
- Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nature Med.* **16**, 793–798 (2010).
- Krzywinski, M. *et al.* Circo: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
- Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
- Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
- Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
- Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
- Pleasant, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
- Pleasant, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
- Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).
- Kwon, J. E. *et al.* BTB domain-containing speckle-type POZ protein (SPOP) serves as an adaptor of Daxx for ubiquitination by Cul3-based ubiquitin ligase. *J. Biol. Chem.* **281**, 12664–12672 (2006).
- Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
- Gaspar-Maia, A. *et al.* Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863–868 (2009).
- Zhang, C. L., McKinsey, T. A. & Olson, E. N. Association of class II histone deacetylases with heterochromatin protein 1: potential role for histone methylation in control of muscle differentiation. *Mol. Cell Biol.* **22**, 7302–7312 (2002).
- Bagchi, A. *et al.* CHD5 is a tumor suppressor at human 1p36. *Cell* **128**, 459–475 (2007).
- Pearl, L. H., Prodromou, C. & Workman, P. The Hsp90 molecular chaperone: an open and shut case for treatment. *Biochem. J.* **410**, 439–453 (2008).
- Kantoff, P. W. *et al.* Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N. Engl. J. Med.* **363**, 411–422 (2010).
- Kantoff, P. W. *et al.* Overall survival analysis of a phase II randomized controlled trial of a Poxviral-based PSA-targeted immunotherapy in metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **28**, 1099–1105 (2010).
- Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
- Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
- Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genet.* **36**, 1065–1071 (2004).
- Lin, C. *et al.* Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139**, 1069–1083 (2009).
- Mani, R. S. *et al.* Induced chromosomal proximity and gene fusions in prostate cancer. *Science* **326**, 1230 (2009).
- Haffner, M. C. *et al.* Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nature Genet.* **42**, 668–675 (2010).
- Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and *TMPRSS2-ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
- Lin, B. *et al.* Integrated expression profiling and ChIP-seq analyses of the growth inhibition response program of the androgen receptor. *PLoS ONE* **4**, e6589 (2009).
- Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
- Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).

39. Krum, S. A. *et al.* Unique ER α cistromes control cell type-specific gene regulation. *Mol. Endocrinol.* **22**, 2393–2406 (2008).
40. Yang, Y., Sterling, J., Storici, F., Resnick, M. A. & Gordenin, D. A. Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. *PLoS Genet.* **4**, e1000264 (2008).
41. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
42. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
43. Wu, X. *et al.* Evidence for regulation of the PTEN tumor suppressor by a membrane-localized multi-PDZ domain containing scaffold protein MAGI-2. *Proc. Natl Acad. Sci. USA* **97**, 4233–4238 (2000).
44. Vazquez, F. *et al.* Phosphorylation of the PTEN tail acts as an inhibitory switch by preventing its recruitment into a protein complex. *J. Biol. Chem.* **276**, 48627–48630 (2001).
45. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
46. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
47. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
48. Carver, B. S. *et al.* Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nature Genet.* **41**, 619–624 (2009).
49. King, J. C. *et al.* Cooperativity of *TMPRSS2-ERG* with PI3-kinase pathway activation in prostate oncogenesis. *Nature Genet.* **41**, 524–526 (2009).
50. Han, B. *et al.* Fluorescence in situ hybridization study shows association of PTEN deletion with ERG rearrangement during prostate cancer progression. *Mod. Pathol.* **22**, 1083–1093 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Leung and all members of the Broad Institute Sequencing Platform. This work was supported by the Prostate Cancer Foundation/Movember (T.R.G., M.A.R., L.A.G.), the Howard Hughes Medical Institute (T.R.G.), the National Human Genome Research Institute (S.B.G., E.S.L.), the Kohlberg Foundation (P.W.K., L.A.G.), the National Cancer Institute (NCI; F.D., M.A.R., M.M.), the NCI Early Detection Research Network (M.A.R.), the National Institutes of Health (L.A.G.), the Department of Defense (F.D.), the Dana-Farber/Harvard Cancer Center Prostate Cancer SPORE grant 2 P50 CA090381-11 (P.W.K.), and the Starr Cancer Consortium (M.F.B., F.D., M.A.R., L.A.G.).

Author Contributions E.S.L., G.G., M.A.R. and L.A.G. contributed equally to this work. M.F.B., E.S.L., G.G., M.A.R., and L.A.G. designed the study, analysed the data, and wrote the paper. M.S.L., F.D., and Y.D. performed analysis of mutations, copy number, rearrangements and ChIP-seq associations. K.C. and A.Y.S. performed analysis of mutations and indels. R.E., D.P., N.K., A.T. and M.A.R. contributed to the procurement of tumour tissue and preparation of DNA. C.S., R.O., W.W., S.M. and K.A. participated in DNA sample processing, quality control, and SNP microarray experiments. L.A., J.W., S.F., J.B. and S.B.G. generated the DNA sequence data. A.S., S.L.C., L.H., T.F., G.S., D.V., A.H.R. and T.J.P. provided additional bioinformatic analyses. K.P., T.Y.M. and M.A.R. performed FISH experiments. M.F.B., M.S.L., R.O., M.P., W.W., G.G. and L.A.G. validated candidate rearrangements. J.W.S., P.W.K., L.C., S.B.G., M.B.G., T.R.G. and M.M. contributed to the study design and interpretation of data.

Author Information All Illumina sequence data have been deposited in dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) and are available at accession phs000330.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike Licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.A.R. (rubinma@med.cornell.edu) or L.A.G. (levi_garraway@dfci.harvard.edu).

Functional identification of an aggression locus in the mouse hypothalamus

Dayu Lin^{1,2}, Maureen P. Boyle³, Piotr Dollar⁴, Hyosang Lee¹, E. S. Lein³, Pietro Perona⁴ & David J. Anderson^{1,2}

Electrical stimulation of certain hypothalamic regions in cats and rodents can elicit attack behaviour, but the exact location of relevant cells within these regions, their requirement for naturally occurring aggression and their relationship to mating circuits have not been clear. Genetic methods for neural circuit manipulation in mice provide a potentially powerful approach to this problem, but brain-stimulation-evoked aggression has never been demonstrated in this species. Here we show that optogenetic, but not electrical, stimulation of neurons in the ventromedial hypothalamus, ventrolateral subdivision (VMHvl) causes male mice to attack both females and inanimate objects, as well as males. Pharmacogenetic silencing of VMHvl reversibly inhibits inter-male aggression. Immediate early gene analysis and single unit recordings from VMHvl during social interactions reveal overlapping but distinct neuronal subpopulations involved in fighting and mating. Neurons activated during attack are inhibited during mating, suggesting a potential neural substrate for competition between these opponent social behaviours.

A central problem in neuroscience is to understand how instinctive behaviours¹, such as aggression, are encoded in the brain. Classic experiments in cats have demonstrated that attack behaviour can be evoked by electrical stimulation of the hypothalamus^{2,3}. However, the precise location of the relevant neurons, and their relationship to circuits for other instinctive social behaviours, such as mating, remain unclear. Studies in the rat have identified a broadly distributed 'hypothalamic attack area' (HAA)^{4–8} that partially overlaps several anatomic nuclei⁹. In contrast, neurons involved in predator defence and mating seem to respect the boundaries of specific, and complementary, hypothalamic nuclei^{10,11}. How aggression circuits are related to these two homologous distinct behavioural subsystems^{9,10} remains poorly understood (but see ref. 12). Immediate early gene (IEG) mapping experiments have suggested that aggression and mating involve similar limbic structures^{13–15}, but whether this reflects the involvement of the same or different cells within these structures is not clear.

We have investigated the localization of hypothalamic neurons involved in aggression, and their relationship to neurons involved in mating, in the male mouse. Using a combination of genetically based functional manipulations and electrophysiological methods, we identify an aggression locus within the ventrolateral subdivision of VMH (VMHvl)⁹. Surprisingly, this structure also contains distinct neurons active during male-female mating. Many neurons activated during aggressive encounters are inhibited during mating. These data indicate a close neuroanatomical relationship between aggression and reproductive circuits, and a potential neural substrate for competition between these social behaviours¹.

Results

Intermingled mating and fighting neurons

We first employed conventional non-isotopic analysis of *c-fos* (also known as *Fos*) induction, a surrogate marker of neuronal excitation¹⁶, to map activity during offensive aggression in the resident-intruder test¹⁷. For comparison, we performed a similar analysis during mating with females. Mating and fighting induced *c-fos* mRNA in the medial amygdala, medial hypothalamus and bed nucleus of the stria terminalis (BNST;

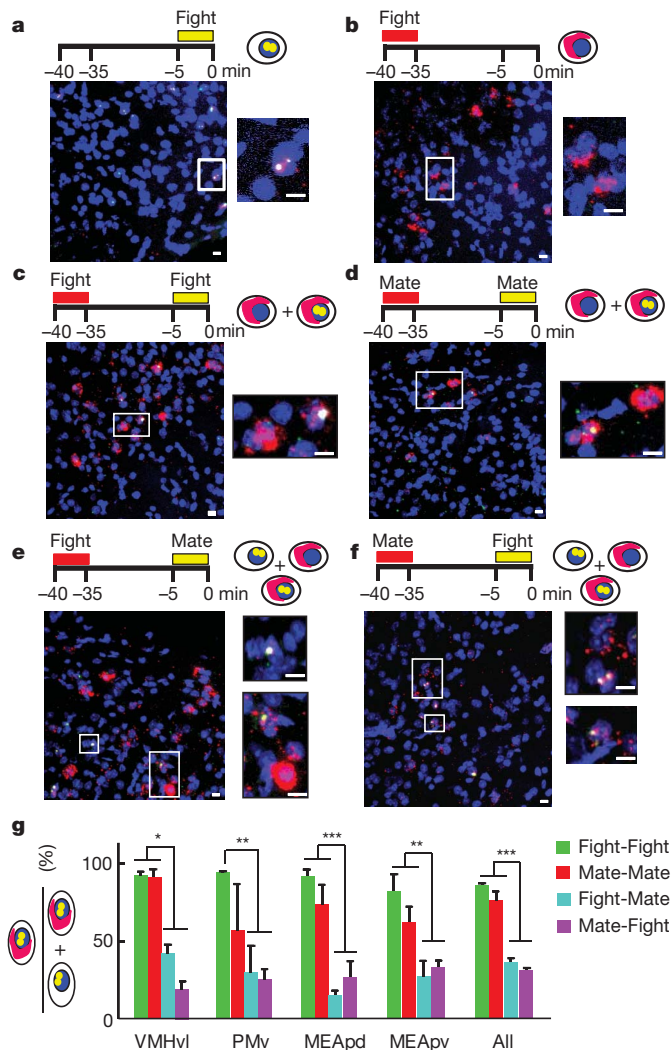
Supplementary Fig. 1), as described previously in rats and hamsters^{13,15}, but not in the anterior hypothalamic nucleus (AHN) which has been implicated in aggression by many studies^{18,19} (reviewed in ref. 20). Whereas the pattern of mating versus fighting-induced *c-fos* was similar in most structures, such between-animal comparisons do not distinguish whether these social behaviours activate the same or different neurons.

To address this issue, we adapted a method, called cellular compartment analysis of temporal activity by fluorescent *in situ* hybridization (catFISH)^{21,22} to compare *c-fos* expression induced during two consecutive behavioural episodes in the same animal (Figs 1a–f). We examined four limbic regions (VMHvl, ventral premammillary nucleus (PMv), medial amygdala posterodorsal (MEApd) and posteroventral (MEApv)) that showed strong *c-fos* induction in single-labelling experiments (Supplementary Fig. 1). Animals killed immediately after 5 min of fighting had almost exclusively nuclear *c-fos* transcripts, whereas those killed 35 min after fighting had essentially only cytoplasmic transcripts (Supplementary Fig. 2). In animals that engaged in two successive episodes of the same behaviour separated by 30 min, most cells expressing nuclear *c-fos* transcripts also expressed cytoplasmic *c-fos* mRNA (Fig. 1c, d, g and Supplementary Fig. 3, green and red bars), indicating activation during both behavioural episodes. By contrast, in animals that sequentially engaged in two different behaviours, only 20–30% of cells with nuclear *c-fos* RNA also expressed cytoplasmic *c-fos* transcripts (Fig. 1e–g and Supplementary Fig. 3, blue and magenta bars). (Nevertheless, the overlap between nuclear and cytoplasmic *c-fos* hybridization was slightly greater than expected by chance even when the two sequential behaviours were different (Supplementary Fig. 4)). These results indicate, first, that the same neurons are likely to be recruited during two successive episodes of mating or fighting, even though such neurons are relatively sparse (Supplementary Fig. 5, <12% of total cells *c-fos*⁺); and second, that mating and fighting may recruit overlapping but distinct sets of neurons in these brain regions.

Chronic recording from the VMHvl

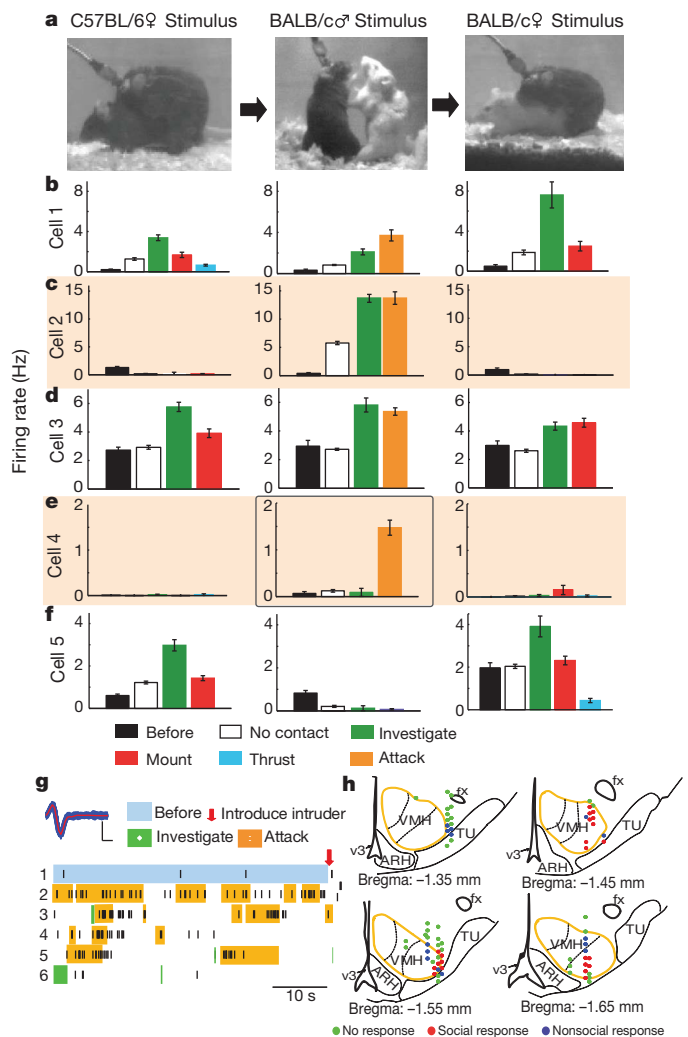
To gain further insight into the relationship between neurons active during mating and fighting, we performed chronic single-unit recordings

¹Division of Biology 216-76, California Institute of Technology, 1201 East California Boulevard, Pasadena, California 91125, USA. ²Howard Hughes Medical Institute, California Institute of Technology, 1201 East California Boulevard, Pasadena, California 91125, USA. ³Allen Institute for Brain Science, Seattle, Washington, 98103, USA. ⁴Computation and Neural Systems 136-93, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.



in awake, behaving male mice using a 16-wire electrode bundle²³ (see Methods). We selected VMHvl for these studies, because it showed preferential *c-fos* induction after fighting versus mating (Supplementary Fig. 5; aggression-induced *c-fos* in VMHvl was further confirmed by double-labelling for *c-fos* and *vglut2*, a glutamate transporter enriched in VMH; Supplementary Fig. 6), and because it overlaps partially with the rat HAA⁷. Recording from VMHvl is challenging because of its deep location, and small size; in only 5 of 30 implanted animals were all 16 electrode tracks confined to VMHvl (Supplementary Fig. 7). Neurons excited during social behaviours (Fig. 2h, red dots) were rarely found among the 25 mistargeted animals. We recorded successfully from 104 well-isolated cells in the five VMHvl-targeted animals. By holding the same cell during alternating, sequential exposures to female and male stimulus animals (Fig. 2a and Supplementary Fig. 8), we could distinguish whether the unit was activated by males and/or females (see Methods for unit isolation criteria).

Neuronal activity patterns in VMHvl during social encounters showed diverse temporal dynamics and sex-selectivity (Figs 2, 3 and Supplementary Figs 8 and 9). Spontaneous firing rates before introduction



of the stimulus animal were typically low (median = 1.1 Hz, range 0–12.7 Hz) and rarely increased during home cage behaviours (that is, grooming); some cells were completely silent until the stimulus animal was presented. Spiking activity was correlated with behaviour by computer-assisted manual annotation of videotape (see Methods). Over 50% (53/104) of recorded cells increased their firing rate during at least one behavioural episode of a social encounter (Fig. 3c). A large fraction (41%; 43/104) of VMHvl cells showed increased firing during an encounter with a male stimulus animal, and on average spiking activity increased with escalation of the encounter, independent of intruder strain (Fig. 3e). In many cases (19/43) this increase began as soon as the intruder male was introduced, and continued as the social encounter progressed, whereas in a comparable number increased firing was observed only during close investigation and subsequent attack (Fig. 2b, d, middle plots, Fig. 3a and Supplementary Movie 1).

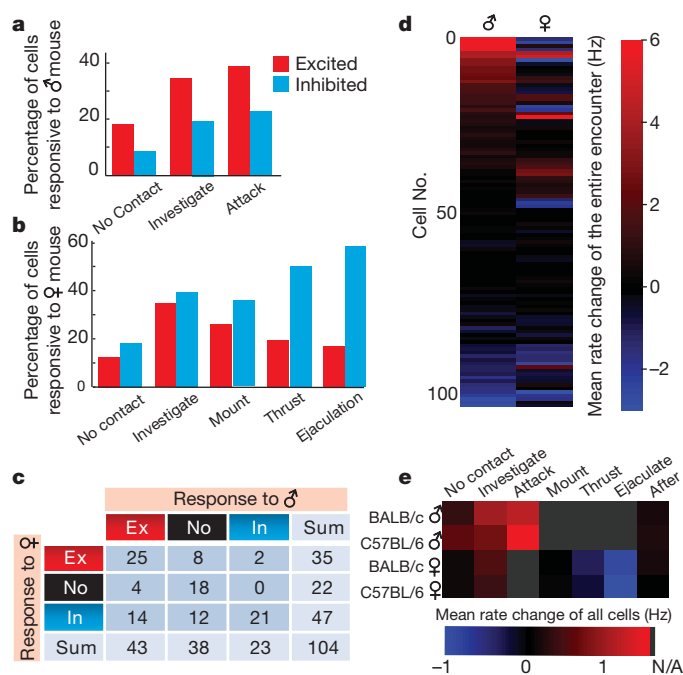


Figure 3 | Summary of cell responses in VMHvl during mating and fighting. **a, b**, Percentage of cells excited (red) or inhibited (blue) during encounters with male (**a**) or female (**b**) mice. **c**, Numbers of cells exhibiting statistically significant changes in firing rate (see Methods) towards males or females. **d**, Firing rate changes for all 104 recorded cells, averaged over entire encounter with males or females. **e**, Firing rate changes averaged over all 104 recorded cells, during various behavioural episodes. Grey, behaviour not applicable (N/A) to the stimulus animal.

Strikingly, a small subset of cells activated in male-male encounters (12%; 5/43) was excited exclusively during attack (Fig. 2e, middle plot, and Fig. 2g).

In contrast, during encounters with females, spiking activity in VMHvl tended to increase only transiently during the initial investigative phase, and subsequently declined as mating progressed (Fig. 3e). Among 35 cells that were excited during female investigation, almost two-thirds (23/35) decreased their firing during subsequent mounting (Fig. 2b and f, left, right, and Fig. 3b), and seven were suppressed (below their baseline firing rate) during thrust and ejaculation (Fig. 2f, right, and Supplementary Fig. 9c, d). Almost half (25/53) of all cells activated during social encounters were excited by both males and females, although most of the largest increases in activity were in sex-specific cells (see Supplementary Footnote 1 and Supplementary Fig. 10). Furthermore, most of this overlap was transitory, occurring during the initial stages of the social encounter and diminishing as the interaction progressed to the consummatory phase of attack or copulation. The observation of partially overlapping populations of male and female excited cells in VMHvl qualitatively confirms the results of our *c-fos* catFISH studies (Supplementary Fig. 10 and Supplementary Footnote 1). However, the evolving segregation of the two populations as the social encounters progressed was not anticipated by the IEG analysis, due to its insufficient temporal resolution.

Our electrophysiological recordings also showed that the majority (14/18) of male excited cells were actively suppressed (below their baseline firing rates) during encounters with females (Figs 2c and 3c, d, Supplementary Fig. 8a, c, e, g and Supplementary Movie 2). Most (86%; 12/14) of those cells, moreover, responded to male intruders before any physical contact. This observation suggests that cells excited during the initiation of an aggressive encounter are selectively suppressed during interactions with a female. In contrast, of the 10 cells selectively excited by females, only two were actively suppressed during

a male-male encounter (Fig. 2f, middle). This asymmetry in sex-specific inhibitory responses indicates that suppression of fighting-related neurons during mating is more pronounced than the converse.

Optogenetic stimulation induced attack

We next tested whether functional manipulations of VMHvl would affect mating or fighting. Although VMHvl overlaps the rat HAA^{7,8,24}, extensive attempts to elicit attack by conventional electrical stimulation of this region in mice were unsuccessful (see Supplementary Footnote 2 and Supplementary Fig. 11). As an alternative, therefore, we expressed channelrhodopsin-2 (ChR2) in VMHvl neurons unilaterally, using stereotactic co-injection of adeno-associated viral vectors (AAV2) expressing Cre recombinase and a Cre-dependent form of ChR2 fused with enhanced yellow fluorescent protein (ChR2-EYFP)^{25,26}, and selectively illuminated cells in this region using an implanted fibre-optic cable²⁷ (Fig. 4a). Because AAV2 infects neurons preferentially²⁸ (Supplementary Fig. 12) and does not retrogradely infect cells from their axons or nerve terminals²⁹, only neurons whose cell bodies are local to the injection site express ChR2 (Supplementary Footnote 3). Optrode recording in anaesthetized animals confirmed that ChR2-expressing cells in VMH can be driven to fire with high temporal precision (Supplementary Fig. 13). Consistent with this result, *c-fos* could be strongly induced in VMHvl on the infected, but not the contralateral control side after repeated blue light stimulation in awake behaving animals (Figs 4b–e).

Optogenetic stimulation of VMHvl in the absence of an intruder did not obviously alter behaviour, except for an occasional increase in exploratory activity. In contrast, in the presence of an intruder, illumination elicited a rapid onset of coordinated and directed attack, often towards the intruder's back (Supplementary Movie 3, see Methods for more detailed behavioural description). Importantly, whereas male mice rarely spontaneously attack females or castrated males, 11/16 ChR2-expressing males exhibited attack towards such intruder animals, within 4–5 s after the onset of illumination (Fig. 4l), over multiple trials (Fig. 4k, Test 1, blue bars). In 9/11 animals, attack was induced during a second test session 1–6 days later (Fig. 4k, Trial 2). Animals with low infection (<10 cells per section, $N = 4$) or animals injected with saline during the surgery ($N = 4$) showed no obvious behavioural changes during light stimulation.

Interestingly, upon illumination offset test animals ceased attack towards females significantly faster than towards castrated males (Fig. 4l, Attack offset). Furthermore, when low intensity (1 mW mm^{-2}) light was used, castrated males were attacked more readily than females (Fig. 4m). We also tested whether illumination could induce attack towards anaesthetized intruders or inanimate objects. Six of 10 animals attacked stationary anaesthetized animals upon illumination; all test animals attacked if the anaesthetized intruders were artificially moved (Fig. 4n). Two of 8 test animals attacked a stationary inflated glove, while 6/8 animals attacked if the glove was moved (Fig. 4n and Supplementary Movie 4).

Histological analysis showed that when the majority of infected cells was located in VMHvl, light stimulation effectively induced attack (red circles in Fig. 4p). In contrast, freezing and flight were observed when VMHdm and VMHc were infected to an equal or greater extent (green circles in Fig. 4p)³⁰. Infection in other regions, such as VMHvl anterior, the lateral hypothalamic area (LHA) and tubular nucleus (TU) was not associated with illumination-induced behavioural changes (Supplementary Figs 14 and 15, Supplementary Footnote 4). To test more directly whether neurons in regions of the HAA⁸ surrounding VMHvl are sufficient to induce aggression, we deliberately infected such regions with AAV2-ChR2. No attack could be induced by light stimulation in such animals ($N = 5$). Strikingly, in cases where the AAV2-ChR2 spread into VMHvl, attack was induced ($N = 3$) (Supplementary Fig. 16). These data indicate that neurons located within VMHvl, but not in adjacent regions, have a key role in mouse aggression.

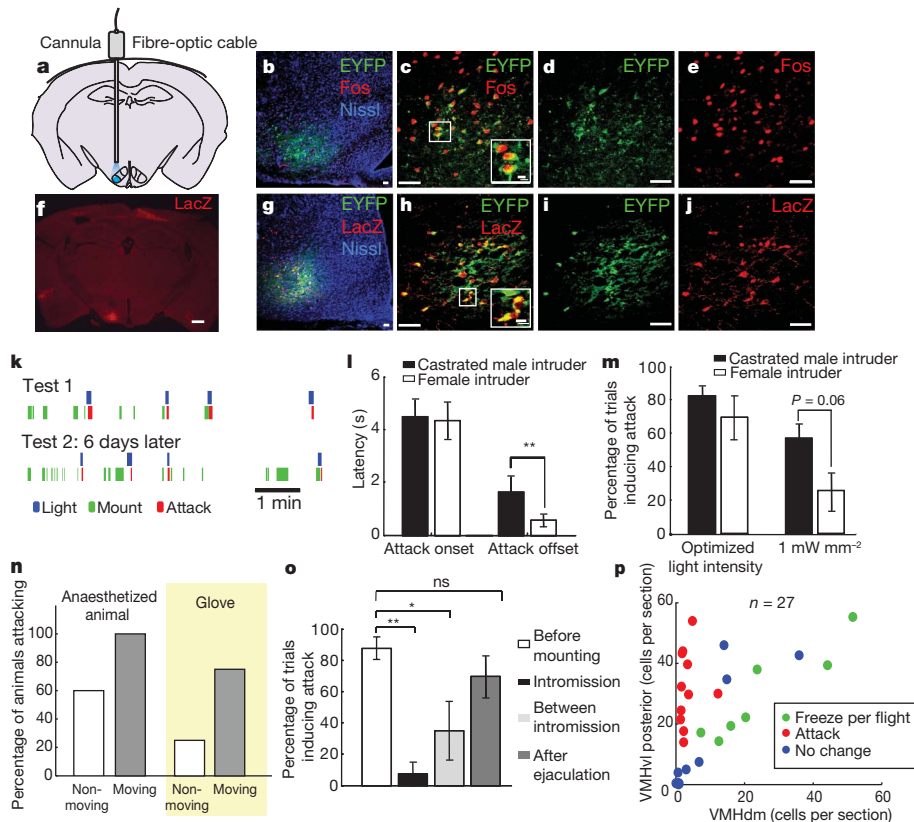


Figure 4 | Optogenetic activation of VMHvl elicits attack in mice.

a, Schematic illustrating optic fibre placement; VMHvl shaded in blue. **b–e**, Fos induction (red) in EF1 α ::ChR2–EYFP-expressing (green) cells at 1 h post-illumination. Fos⁺ cells outside EYFP⁺ region may be synaptic targets of ChR2-activated cells. Blue, fluorescent Nissl stain. **f**, LacZ expression identifies infected cell bodies (red). Scale bar, 500 μ m. **g–j**, LacZ expression (red) and native ChR2–EYFP fluorescence (green) largely overlaps. Boxed areas in **c**, **h** enlarged at lower right. Scale bars in **b–e**, **g–j**, 50 μ m or 10 μ m (insets). **k**, Raster plots illustrating behavioural episodes (legend below) in a ChR2-expressing male paired with a female in two consecutive tests. **l**, Attack onset/offset latencies (relative to initiation versus termination of illumination) towards indicated intruders, $^{**}P < 0.01$. **m**, Efficacy of light-stimulated attack.

The observations that overall activity in VMHvl decreases during male–female mating (Fig. 3e), and that many male excited cells are inhibited by females (Fig. 3d), indicated that a progressive inhibition of attack neurons occurs as mating progresses towards its consummatory phase. To test this, we stimulated VMHvl during encounters with females, before mounting, during intromission, between intromissions and after ejaculation. When illumination was delivered before mounting, attack towards the female was elicited in over 80% of trials at light intensities between 1 and 2 mW mm^{−2}, in all seven tested animals (Fig. 4o, white bar). But during intromission, the same light intensity was often ineffective, even with extended stimulation (Fig. 4o, black bar). Increasing the light intensity fourfold elicited female-direct attack during intromission in five of seven animals, but with increased latency (Supplementary Fig. 17). Between intromissions, attack was evoked in 30% of cases. Strikingly, following ejaculation the frequency of illumination-evoked attack recovered to pre-mounting levels (Fig. 4o, dark grey bar). Thus mating exerts an increasingly strong suppression of optogenetically stimulated attack, as the encounter progresses towards its consummatory phase.

Mouse aggression requires VMHvl activity

Whether neurons that mediate brain-stimulation-evoked attack are also required for naturally occurring aggression has been controversial. Electrolytic lesions of VMH in rats and mice have yielded seemingly contradictory results^{31,32}, and this method destroys axons-of-passage

‘Optimized light intensity’, laser power yielding average maximal response in each animal (range: 1–3.3 mW mm^{−2}). ‘1 mW mm^{−2}’, average response obtained at this power (*t*-test, $P = 0.06$). **n**, Percentage of animals attacking moving versus non-moving anaesthetized animals or inflated glove (yellow shading). **o**, Percentage of trials inducing attacks towards female during successive stages of mating. $^{*}P < 0.05$, $^{**}P < 0.01$ (one-way ANOVA with Bonferroni correction). ns, not significant. **p**, Distribution of infected cells in each animal, plotted as cells per section in VMHvl posterior portion versus that in (VMHdm + VMHc) region. Colour code indicates whether illumination induced freeze/flight (green), attack (red) or no change in behaviour (blue). See also Supplementary Footnote 4 for further statistical analysis.

as well as cell bodies. There is little evidence that local chemical inhibition of neuronal activity in the rat HAA reduces aggression (although inhibition³³ or killing³⁴ of Substance P receptor-expressing neurons attenuates ‘hard biting’ behaviour). We therefore asked whether reversible genetic suppression of electrical excitability in VMHvl neurons inhibits attack behaviour. To do this, we used separate AAV2 vectors to co-express two subunits (α and β) comprising a *Caenorhabditis elegans* ivermectin (IVM)-gated chloride channel (GluCl $\alpha\beta$)³⁵, which has been mutated to eliminate glutamate sensitivity³⁶. Upon IVM binding, this heteropentameric channel prevents action potential firing by hyperpolarizing the membrane^{28,35}.

Three weeks after viral injection, animals were administered IVM intraperitoneally 24 h before testing²⁸. The experimental group ($N = 33$) showed a decrease in the total attack duration, and an increase in the latency to the first attack, in comparison to saline-injected or GluCl β -only injected controls (Figs 5f, g; see Methods). Furthermore, 25% of the experimental animals failed to initiate any attack during the post-IVM test. Experimental animals performed similarly in the rotarod assay before and after IVM administration, indicating no change in motor coordination or fatigue (Supplementary Fig. 18). Eight days after the IVM injection, the aggression level of the test group recovered to the pre-IVM level and could be suppressed again by a second IVM injection (Fig. 5h). Immunohistochemical analysis (Fig. 5a, d) indicated a reverse correlation between the suppression of aggression and the percentage of GluCl-expressing cells in

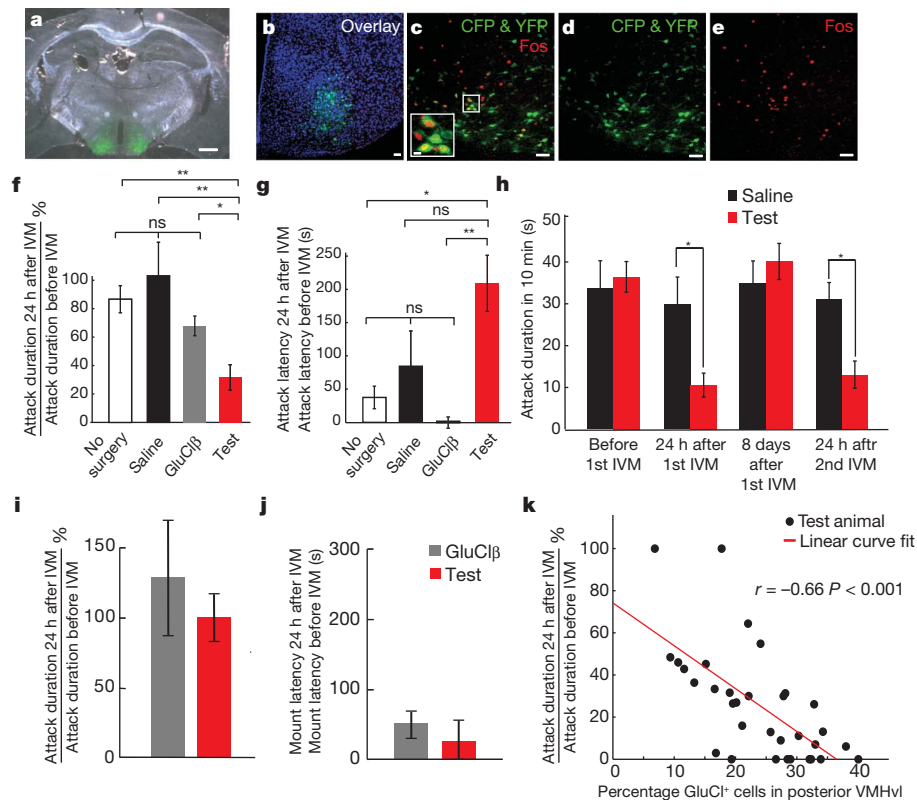


Figure 5 | Reversible inhibition of natural aggression by genetic silencing of VMHvl **a**, Anti-GFP antibody staining (green) in mice bilaterally infected with AAV2-GluCl α + AAV2-GluCl β . Scale bar, 500 μ m. **b–e**, Overlap between GluCl-expressing (green) and Fos-expressing (red) cells, 1 h after fighting. Blue, Topro-3 nuclear stain. Inset in **c** represents boxed area. Yellow cells are double-labelled. Scale bars, 50 μ m or 10 μ m (inset). **f**, **g**, Percent change in cumulative attack duration (**f**) and latency (**g**) during a 600 s resident-intruder trial before versus 24 h after IVM injection. Test, GluCl virus-injected animals ($n = 33$) (red bar). Control, no surgery (white bar, $n = 12$), saline (black bar, $n = 6$) or GluCl β virus-injected animals ($n = 12$, grey bar) (** $P < 0.01$, * $P < 0.05$,

t-test). **h**, Cumulative attack duration during repeated cycles of IVM injection or washout (* $P < 0.05$, Bonferroni after tests of two-way ANOVA with repeated measures). Test, GluCl virus-injected animals ($n = 12$); Control, saline ($n = 6$). **i**, **j**, Percent change in mount duration (**i**) or latency (**j**) in test ($n = 12$) versus control (GluCl β virus injected, $n = 12$) males paired with females. **k**, Percentage of infected cells in posterior portion of VMHvl (Bregma -1.4 – 1.8 mm) plotted against extent of aggression suppression after IVM injection. The Pearson correlation coefficient is significantly higher than 0 ($P < 0.001$). See Supplementary Fig. 19 for further analysis.

the posterior half of VMHvl (Bregma level -1.4 – 1.75 mm; Fig. 5k). No such correlation was found in VMHvl anterior (Bregma level -1.15 – 1.4 mm) or in other regions surrounding VMHvl (Supplementary Fig. 19). Double-label immunostaining for GFP and Fos in animals killed 1 h after an aggressive interaction (following IVM washout) indicated that viral infection overlapped the population activated during fighting ((GFP⁺ Fos⁺)/total Fos⁺ > 50%; $n = 4$; Fig. 5b–e). These data indicate that genetic silencing of neurons in VMHvl can reversibly inhibit aggressive behaviour. In GluCl-expressing males paired with females, no change in mounting duration or latency to the first mount was observed after IVM injection (Figs 5i, j). Because the overall level of neuronal activity in VMHvl is normally suppressed during the consummatory phase of mating (Fig. 3e), it is not surprising that further inhibition of activity failed to impair such behaviour.

Discussion

Using genetically based manipulations in mice, we show that neurons necessary and sufficient for offensive aggression are localized within a small subdivision of VMH. The more diffuse HAA identified in rats^{6,24} may reflect a species difference, or the fact that electrical stimulation mapping³⁷ activates both axons-of-passage and neuronal somata, whereas our manipulations are restricted to the latter. Our *in vivo* recordings indicate that some neurons in VMHvl are activated by intruder conspecifics before physical contact. This suggests a function in olfactory coding, perhaps related to sex discrimination. However, optogenetic stimulation of VMHvl evoked aggressive behaviour

towards an inanimate object, arguing for a causal role in the motivation or drive to attack. We suggest that VMHvl has a key role in sensori-motor transformations and/or the encoding of motivational states⁹ underlying aggression. The relationship of the aggression circuits within VMHvl to those involved in defensive^{12,20,38} or maternal^{139,40} aggression remains to be investigated.

Whereas VMH is well established to have a key role in female reproductive behaviour^{41,42}, it has not traditionally been considered as a key node in male mating circuitry¹¹ (but see ref. 13). We have identified cells within the VMHvl of males that are activated during male-female mating, and which are mostly distinct from those activated during fighting. The role of these neurons is not yet clear, because our functional manipulations did not perturb mating behaviour. One possibility is that these female-activated neurons serve to inhibit aggression during mating. Consistent with this idea, many male-activated units were actively inhibited by females, and a higher intensity of illumination was required to evoke attack towards a female during mating encounters. These data identify a neural correlate of competitive interactions between fighting and mating¹. Whether this competition originates in VMHvl, or is controlled by descending inputs to this nucleus⁴³, awaits further investigation.

METHODS SUMMARY

Sexually experienced C57BL/6N male mice, singly housed on a reverse light-dark cycle, were used. Resident-intruder assays were designed to maximize offensive aggression by the resident; no attacks were initiated by the intruder under our conditions. For *in situ* hybridization, animals were killed 30 min after a 10 min

standard resident-intruder assay and processed as described⁴⁴. For Fos catFISH experiments, animals experienced two 5-min behavioural episodes 30 min apart, and were killed immediately after the second episode. An intronic *c-fos* probe and a *c-fos* cRNA probe were combined to detect nuclear *c-fos* primary transcripts. For chronic recording, a movable bundle of sixteen 13- μ m tungsten microwires was implanted, and 2 weeks allowed for recovery. On recording days, a flexible cable was attached to the microdrive and connected to a commutator. Recordings were performed in the animals' home cage. Female and male mice were introduced for approximately 10 min per session. Spiking activity and behaviour were synchronously recorded. Data analysis, including behavioural annotation of videotapes, was performed using custom software written in Matlab. For Chr2 experiments, 150 nl of a 4:2:1 mixture of an AAV2 Cre inducible EF1 α ::Chr2 (ref. 26), AAV2 CMV::CRE and AAV2 CMV::LacZ with a similar final titre (8×10^{11} p.f.u. ml⁻¹) was stereotactically injected unilaterally. After 2 weeks of recovery, light pulses (20 ms, 20 Hz, 1–4 mW mm⁻²) were delivered to activate the targeted region for 2–20 s in the presence of various stimuli. For GluCl inactivation experiments, animals in the experimental group (bilateral injection of 10^9 viral particles per side of AAV2 expressing cyan-fluorescent-protein- and yellow-fluorescent-protein-tagged GluCl α and GluCl β , respectively, under the control of the CAG promoter (CAG::GluCl α -CFP and CAG::GluCl β -YFP)), and each of the three control groups (no surgery, saline or GluCl β bilaterally injected) were tested three times in the resident-intruder assay to establish a stable aggression baseline. After the third test, IVM (1%, 5 mg kg⁻¹) was injected intraperitoneally and the animals were tested again 24 h and 8 days later.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 July; accepted 2 December 2010.

1. Tinbergen, N. *The study of instinct* (Clarendon Press/Oxford University Press, 1951).
2. Hess, W. R. Stammganglien-Reizversuche. Berichte der gesamten. *Physiologie* **42**, 554–555 (1928).
3. Hess, W. R. & Brügger, M. Das subkortikale Zentrum der affektiven Abwehrreaktion. *Helv. Physiol. Acta* **1**, 33–52 (1943).
4. Hrabovszky, E. et al. Neurochemical characterization of hypothalamic neurons involved in attack behavior: glutamatergic dominance and co-expression of thyrotropin-releasing hormone in a subset of glutamatergic neurons. *Neuroscience* **133**, 657–666 (2005).
5. Kruk, M. R. et al. Discriminant analysis of the localization of aggression-inducing electrode placements in the hypothalamus of male rats. *Brain Res.* **260**, 61–79 (1983).
6. Kruk, M. R. Ethology and pharmacology of hypothalamic aggression in the rat. *Neurosci. Biobehav. Rev.* **15**, 527–538 (1991).
7. Lammers, J. H., Kruk, M. R., Meelis, W. & van der Poel, A. M. Hypothalamic substrates for brain stimulation-induced attack, teeth-chattering and social grooming in the rat. *Brain Res.* **449**, 311–327 (1988).
8. Siegel, A., Roeling, T. A., Gregg, T. R. & Kruk, M. R. Neuropharmacology of brain-stimulation-evoked aggression. *Neurosci. Biobehav. Rev.* **23**, 359–389 (1999).
9. Swanson, L. W. Cerebral hemisphere regulation of motivated behavior. *Brain Res.* **886**, 113–164 (2000).
10. Canteras, N. S. The medial hypothalamic defensive system: hodological organization and functional implications. *Pharmacol. Biochem. Behav.* **71**, 481–491 (2002).
11. Simerly, R. B. Wired for reproduction: organization and development of sexually dimorphic circuits in the mammalian forebrain. *Annu. Rev. Neurosci.* **25**, 507–536 (2002).
12. Motta, S. C. et al. Dissecting the brain's fear system reveals the hypothalamus is critical for responding in subordinate conspecific intruders. *Proc. Natl Acad. Sci. USA* **106**, 4870–4875 (2009).
13. Kollack-Walker, S. & Newman, S. W. Mating and agonistic behavior produce different patterns of Fos immunolabeling in the male Syrian hamster brain. *Neuroscience* **66**, 721–736 (1995).
14. Newman, S. W. The medial extended amygdala in male reproductive behavior. A node in the mammalian social behavior network. *Ann. NY Acad. Sci.* **877**, 242–257 (1999).
15. Veening, J. G. et al. Do similar neural systems subserve aggressive and sexual behaviour in male rats? Insights from c-Fos and pharmacological studies. *Eur. J. Pharmacol.* **526**, 226–239 (2005).
16. Morgan, J. I., Cohen, D. R., Hempstead, J. L. & Curran, T. Mapping patterns of c-fos expression in the central nervous system after seizure. *Science* **237**, 192–197 (1987).
17. Blanchard, D. C. & Blanchard, R. J. Ethoexperimental approaches to the biology of emotion. *Annu. Rev. Psychol.* **39**, 43–68 (1988).
18. Delville, Y., De Vries, G. J. & Ferris, C. F. Neural connections of the anterior hypothalamus and agonistic behavior in golden hamsters. *Brain Behav. Evol.* **55**, 53–76 (2000).
19. Ferris, C. F. & Potegal, M. Vasopressin receptor blockade in the anterior hypothalamus suppresses aggression in hamsters. *Physiol. Behav.* **44**, 235–239 (1988).
20. Nelson, R. J. & Trainor, B. C. Neural mechanisms of aggression. *Natl. Rev.* **8**, 536–546 (2007).
21. Guzowski, J. F., McNaughton, B. L., Barnes, C. A. & Worley, P. F. Imaging neural activity with temporal and cellular resolution using FISH. *Curr. Opin. Neurobiol.* **11**, 579–584 (2001).
22. Guzowski, J. F., McNaughton, B. L., Barnes, C. A. & Worley, P. F. Environment-specific expression of the immediate-early gene *Arc* in hippocampal neuronal ensembles. *Nature Neurosci.* **2**, 1120–1124 (1999).
23. Herry, C. et al. Switching on and off fear by distinct neuronal circuits. *Nature* **454**, 600–606 (2008).
24. Roeling, T. A. et al. Efferent connections of the hypothalamic “aggression area” in the rat. *Neuroscience* **59**, 1001–1024 (1994).
25. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8**, 1263–1268 (2005).
26. Kravitz, A. V. et al. Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature* **466**, 622–626 (2010).
27. Aravanis, A. M. et al. An optical neural interface: *in vivo* control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.* **4**, S143–S156 (2007).
28. Lerchner, W. et al. Reversible silencing of neuronal excitability in behaving mice by a genetically targeted, ivermectin-gated Cl⁻ channel. *Neuron* **54**, 35–49 (2007).
29. Taymans, J. M. et al. Comparative analysis of adeno-associated viral vector serotypes 1, 2, 5, 7, and 8 in mouse brain. *Hum. Gene Ther.* **18**, 195–206 (2007).
30. Lammers, J. H., Kruk, M. R., Meelis, W. & van der Poel, A. M. Hypothalamic substrates for brain stimulation-induced patterns of locomotion and escape jumps in the rat. *Brain Res.* **449**, 294–310 (1988).
31. Olivier, B. Ventromedial hypothalamus and aggressive behavior in rats. *Aggress. Behav.* **3**, 47–56 (1977).
32. Olivier, B. & Wiepkema, P. R. Behaviour changes in mice following electrolytic lesions in the median hypothalamus. *Brain Res.* **65**, 521–524 (1974).
33. Halasz, J. et al. The effect of neurokinin1 receptor blockade on territorial aggression and in a model of violent aggression. *Biol. Psychiatry* **63**, 271–278 (2008).
34. Halasz, J. et al. Substance P neurotransmission and violent aggression: the role of tachykinin NK(1) receptors in the hypothalamic attack area. *Eur. J. Pharmacol.* **611**, 35–43 (2009).
35. Slimko, E. M., McKinney, S., Anderson, D. J., Davidson, N. & Lester, H. A. Selective electrical silencing of mammalian neurons *in vitro* by the use of invertebrate ligand-gated chloride channels. *J. Neurosci.* **22**, 7373–7379 (2002).
36. Li, P., Slimko, E. M. & Lester, H. A. Selective elimination of glutamate activation and introduction of fluorescent proteins into a *Caenorhabditis elegans* chloride channel. *FEBS Lett.* **528**, 77–82 (2002).
37. van der Poel, A. M. et al. A locked, non-rotating, completely embedded, moveable electrode for chronic brain stimulation studies in freely moving, fighting rats. *Physiol. Behav.* **31**, 259–263 (1983).
38. Blanchard, R. J., Wall, P. M. & Blanchard, D. C. Problems in the study of rodent aggression. *Horm. Behav.* **44**, 161–170 (2003).
39. Lonstein, J. S. & Gammie, S. C. Sensory, hormonal, and neural control of maternal aggression in laboratory rodents. *Neurosci. Biobehav. Rev.* **26**, 869–888 (2002).
40. Kruk, M. R. et al. Comparison of aggressive behaviour induced by electrical stimulation in the hypothalamus of male and female rats. *Prog. Brain Res.* **61**, 303–314 (1984).
41. Pfaff, D. W. & Sakuma, Y. Facilitation of the lordosis reflex of female rats from the ventromedial nucleus of the hypothalamus. *J. Physiol. (Lond.)* **288**, 189–202 (1979).
42. Pfaff, D. W. & Sakuma, Y. Deficit in the lordosis reflex of female rats caused by lesions in the ventromedial nucleus of the hypothalamus. *J. Physiol. (Lond.)* **288**, 203–210 (1979).
43. Petrovich, G. D., Canteras, N. S. & Swanson, L. W. Combinatorial amygdalar inputs to hippocampal domains and hypothalamic behavior systems. *Brain Res. Rev.* **38**, 247–289 (2001).
44. Mongeau, R., Miller, G. A., Chiang, E. & Anderson, D. J. Neural correlates of competing fear behaviors evoked by an innately aversive stimulus. *J. Neurosci.* **23**, 3855–3868 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Ciochi, S. Lin, Y. Ben-Shaul and A. Wang for advice on electrode and microdrive design; M. Gerfen and M. Vondrus for microdrive fabrication; M. P. Walsh, T. D. Heitzman and V. Rush for electronics support; A. Steele, R. Robbins, S. Ossorio, K. Gunapala, A. Paul, D. Oh, C. Kim and J. Nishiguchi for behaviour annotation and video scoring; H. Kim for technical assistance; W. E. Haubensack for teaching fibre-optic implant methods; J. T. Henderson for advice on stereotactic surgery; M. Kruk for advice on electrical stimulation experiments; K. Deisseroth for the Cre-dependent Chr2–EYFP AAV construct. M. Kruk, R. Mooney and R. Simerly for comments on the manuscripts; G. Mosconi for lab management and G. Mancuso for administrative support. This work was supported in part by the Weston-Havens Foundation and Jane Coffin Childs Foundation. D.J.A. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions D.L. designed, carried out and analyzed preliminary fos catFISH experiments and all other experiments, and co-wrote the manuscript; M.B. and E.L. performed additional fos catFISH experiments; P.D. and P.P. developed custom behaviour annotation software; H.L. performed some of the optogenetic experiments; D.J.A. conceived the project, suggested experiments, analysed data and co-wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.L. (dayu.lin@nyumc.org) and D.J.A. (wuwei@caltech.edu).

METHODS

Behavioural tests. All test animals used in this study were adult proven breeder C57BL/6 male mice (Charles River Laboratory). They were singly housed under a reversed light-dark cycle for at least 1 week before the test. The care and experimental manipulation of the animals were carried out in accordance with the NIH guidelines and approved by the Caltech Institutional Animal Care and Use Committee. For resident-intruder assays, C57BL/6 males were allowed to interact with BALB/c males for 10 min. All intruder mice were group housed, and had similar body weight as the test mice. All resident animals included in the study initiated all the attacks and showed no submissive postures during the aggression test. For mating tests, the residents were allowed to interact with sexually receptive BALB/c and C57BL/6 females for 10 min. Females were screened for receptivity by pairing with a singly housed C57BL/6 male mouse briefly before each test.

In situ hybridization. Brains from mice killed 30 min after performing either the 10 min resident-intruder or mating tests were analysed for expression of *c-fos* mRNA throughout the forebrain, using non isotopic *in situ* hybridization on 120 µm thick sections. Details of the procedure have been described previously⁴⁴. For *fos* catFISH experiments, animals experienced two consecutive 5 min fighting or mating episodes 30 min apart, and were killed immediately after the second episode. Since *c-fos* transcripts are detected only in the nucleus within the first 5 min following induction, and are completely translocated to the cytoplasm as processed mRNA after 35 min, the sub-cellular location of *c-fos* allows one to distinguish neurons activated by a single stimulus from those successively activated by both stimuli: only in the latter case will transcripts be present in both the nucleus and cytoplasm. We used an intronic *fos* probe with a different fluorescent colour label, in addition to the *fos* cRNA probe, which allowed us to more easily differentiate nuclear from cytoplasmic FISH signals. The *c-fos* transcript distribution pattern was examined using both colour combinations (*fos* cRNA probe in green and *fos* intronic probe in red, or vice versa) and no difference was observed. See Supplementary Methods for more detailed dFISH procedure and microscopic analysis.

Electrophysiological recording. A bundle of sixteen tungsten microwires (13 µm diameter each, California Fine Wire) attached to a mechanical microdrive was implanted in one hemisphere and secured with bone screws and dental acrylic during stereotaxic surgery. The drive was a miniaturized version of an original design described elsewhere⁴⁵. Two weeks after initial implantation, and on days of recording, a flexible cable was attached to the microdrive and connected to a torqueless, feedback-controlled commutator (Tucker Davis Technology). During recording sessions, the test animals were allowed to stay in their home cage and interact with the stimulus animals freely. Female or male mice were introduced into the test arena for approximately 10 min. A given type of stimulus (for example, a male mouse) was presented on multiple occasions, to examine the reproducibility of a response. All recordings were carried out in subdued light with infrared illumination. A commercial recording system was used for data acquisition (Tucker Davis Technology). Digital infrared video recordings of animal behaviour from both side and top view were simultaneously streamed to a hard disk at 640 × 480 pixel resolution at 25 frames per second (Streampix, Norpix). Each video frame acquisition was triggered by a TTL pulse from the recording setup to achieve synchronization between the video and the electrophysiological recording. Spikes of individual neurons were sorted using commercial software (OpenSorter, Tuck Davis Technology), based on principal component analysis. Unit isolation was verified using autocorrelation histograms. To ensure that single units were isolated, and that the same units were recorded in the presence of sequentially presented male or female stimulus animals, we imposed four criteria to select cells for subsequent statistical analysis. First, the cells had to have a signal/noise ratio >3; second, the spike shape had to be stable throughout the recording; third, the response had to be repeatable during multiple trials; fourth, the percentage of spikes occurring with inter-spike intervals (ISIs) <3 ms (the typical refractory period for a neuron) in a continuous recording sequence had to be <0.1%. Of the cells included in the analysis, 74 out of 104 had all of their ISIs ≥ 3 ms. After each recording session, the microwire bundle was advanced 70 µm, by adjusting a 00-90 screw on the drive by a quarter of a turn. After 5 to 10 recording sessions, which typically were distributed over 2 to 3 months, animals were euthanized and the location of the recording electrodes verified histologically.

Behavioural annotation and statistical analysis of firing rate changes. Custom software written in Matlab was used to facilitate manual annotation of mouse behaviour from videotaped recording sessions. Annotations were made using side- and top-view videos played simultaneously. A total of ~1,000 10 min videos were carefully analysed on a frame-by-frame basis. The behavioural results were then correlated with the electrophysiology to obtain histograms of firing rates during various behavioural episodes. Firing rates for each unit were averaged in 0.5 s bins, and the mean firing rate during each behavioural episode (for example, 'Investigation') was compared to the baseline firing rate (that is, before introduction of the test animals) using Kruskal–Wallis one-way analysis of variance by

ranks (with *P* value 0.01), followed by a pairwise test for significance with Tukey–Kramer correction for multiple comparisons, to determine whether there was any statistically significant change in activity during a given episode. If the same stimulus was tested multiple times, only repeatable responses were regarded as positive.

ChR2 viral activation. The Cre-inducible EF1α::ChR2–EYFP construct was the gift of K. Deisseroth and was described earlier⁴⁶. Because ChR2 is a membrane protein expressed mainly in axons, we co-injected an AAV2 CMV::LacZ virus to facilitate the quantification and anatomic localization of infected cells. AAV2 CMV::CRE and AAV2 CMV::LacZ viruses were purchased from Vector Biolabs. AAV2 CRE inducible EF1α::ChR2–EYFP virus was prepared by the Harvard Vector Core Facility. The AAV2-ChR2, AAV2-CRE and AAV2-LacZ viruses were mixed in a 4:2:1 volume ratio to reach a similar final titer (8×10^{11} pfu/mL). A total of 0.15 µl of the mixed virus suspension (approximately 1.2×10^8 particles) was injected unilaterally over a period of 5 min using a fine glass capillary (Nanoject II, Drummond Scientific). After injection, a 24 gauge cannula (Plastics One) was inserted and secured to a depth of approximately 0.6 mm above the target region (Metabond, Parkell). After 2 weeks recovery, and on test days, a 200 µm multimode optical fibre (Thorlabs) was inserted into the cannula and secured with an internal cannula adaptor and a cap (Plastics One). The tip of the fibre was cut flat to the bottom of the implanted cannula. Blue (473 nm) light was delivered in 20 ms pulses at 20 Hz, at final output powers ranging from 1 to 4 mW mm⁻² (Crystallaser). Each light stimulation episode lasted from 2 to 20 s, depending on the behavioural responses. Initial tests using various frequencies indicated that 10–15 Hz was the minimal frequency necessary to induce a behavioural response. All animals were tested twice with 1 to 6 days between tests. Light-induced attack typically includes the following steps: the stimulated animal approaches the intruder from a distance, bites the intruder's back repeatedly, then either stops abruptly upon the cessation of light stimulation and moves away, or stops biting gradually after several rounds of attack. Light-induced escape behaviour typically includes the following steps: the stimulated animal makes a quick movement towards the corner of the arena. If the animal is engaged in other behaviours such as fighting or mating, it stops those behaviours and moves to a corner of the cage. Typically, the animal will stay in the corner and maintain the same posture for the remainder of the stimulation period.

One hour before sacrificing the animal, a train of light (10 s on and 10 s off, 20 ms, 20 Hz × 20) was delivered to induce Fos expression in the absence of any target animal. A total of 28 animals were implanted and tested. Twenty seven animals were processed for histological analysis and were included in the scatter plot of Fig. 4p. To quantify the extent of infection, we counted all the LacZ⁺ cells in various regions and calculated the number of LacZ⁺ cells in VMHvl posterior, VMHvl anterior, VMHdm + VMHc, LH and TU for each section. Fluorescent Nissl or NeuN staining was used to determine the boundaries of different VMH subdivisions. In cases where the boundary was hard to determine precisely, we delineated VMHvl as extending from the ventral pole of VMH approximately 1/3 of the way along the dorso-ventral and medio-lateral axes.

GluCl viral inactivation. Animals in the experimental group (*n* = 33) were stereotactically injected bilaterally with a total of 0.9 µl AAV2-GluClα and AAV2-GluClβ, each under the control of the CAG promoter-enhancer, in a 1:1 mixture (approximately 10^9 particles), using a glass capillary attached to an auto nanolitre injector (Drummond). The viral constructs have been described previously²⁸. One control group received no surgery (*n* = 12), a second and third control group received either saline (*n* = 6) or AAV2-GluClβ (*n* = 12) during the surgery. After 2 weeks of recovery, the aggression level of the animal was evaluated using a 10 min resident-intruder assay three times on different days. After the third test, a 1% sterile solution of Ivermectin (Phenectin, AmTech) was injected intraperitoneally at 5 mg kg⁻¹ animal body weight. The animal were then tested again 24 h and 8 days later. The effect of IVM typically wears off completely by 8 days²⁸ and any behavioural change is expected to be reversed at that time point. The mating test was performed using a similar procedure, except that a receptive female mouse was used as the stimulus animal. The rotarod assay was performed as described previously⁴⁷. The animal was exposed to a 10 min resident-intruder assay 1 h before sacrifice to induce Fos expression. The brains were then harvested for histological analysis.

45. Bragin, A. et al. Multiple site silicon-based probes for chronic recordings in freely moving rats: implantation, recording and histological verification. *J. Neurosci. Methods* **98**, 77–82 (2000).
46. Gradinaru, V., Mogri, M., Thompson, K. R., Henderson, J. M. & Deisseroth, K. Optical deconstruction of parkinsonian neural circuitry. *Science* **324**, 354–359 (2009).
47. Southwell, A. L., Ko, J. & Patterson, P. H. Intrabody gene therapy ameliorates motor, cognitive, and neuropathological symptoms in multiple mouse models of Huntington's disease. *J. Neurosci.* **29**, 13589–13602 (2009).

Structural basis for the subunit assembly of the anaphase-promoting complex

Anne Schreiber¹, Florian Stengel², Ziguo Zhang¹, Radoslav I. Enchev¹, Eric H. Kong¹, Edward P. Morris¹, Carol V. Robinson², Paula C. A. da Fonseca¹ & David Barford¹

The anaphase-promoting complex or cyclosome (APC/C) is an unusually large E3 ubiquitin ligase responsible for regulating defined cell cycle transitions. Information on how its 13 constituent proteins are assembled, and how they interact with co-activators, substrates and regulatory proteins is limited. Here, we describe a recombinant expression system that allows the reconstitution of holo APC/C and its sub-complexes that, when combined with electron microscopy, mass spectrometry and docking of crystallographic and homology-derived coordinates, provides a precise definition of the organization and structure of all essential APC/C subunits, resulting in a pseudo-atomic model for 70% of the APC/C. A lattice-like appearance of the APC/C is generated by multiple repeat motifs of most APC/C subunits. Three conserved tetratricopeptide repeat (TPR) subunits (Cdc16, Cdc23 and Cdc27) share related superhelical homo-dimeric architectures that assemble to generate a quasi-symmetrical structure. Our structure explains how this TPR sub-complex, together with additional scaffolding subunits (Apc1, Apc4 and Apc5), coordinate the juxtaposition of the catalytic and substrate recognition module (Apc2, Apc11 and Apc10 (also known as Doc1)), and TPR-phosphorylation sites, relative to co-activator, regulatory proteins and substrates.

The APC/C is an unusually complex multi-subunit E3 ubiquitin ligase endowed with elaborate regulatory, catalytic and specificity properties. By mediating the ubiquitylation of a diverse array of mitotic regulatory proteins, the APC/C controls the cell cycle processes responsible for chromatid segregation at the metaphase to anaphase transition, the completion of mitosis, and the establishment and maintenance of G1 (refs 1–3). Short sequence motifs, predominantly D-box⁴ and KEN-box⁵, target APC/C substrates for ubiquitylation and subsequent degradation by the 26S proteasome. APC/C activation⁶ and substrate recruitment^{7,8} requires interaction with co-activator (either Cdc20 or Cdh1). Phosphorylation of APC/C subunits and co-activators, and interactions with regulatory proteins, controls its activity^{1–3}.

The APC/C is assembled from 13 different proteins (Supplementary Table 1). Larger APC/C subunits incorporate various multiple repeat motifs, including the TPR motifs of Cdc16, Cdc23, Cdc27 (refs 9, 10) and Apc5 (ref. 1), proteasome-cyclosome (PC) repeats of Apc1 (ref. 11), and cullin repeats of Apc2. Apc2 binds the RING domain subunit Apc11 forming the APC/C catalytic core that recruits the E2 ubiquitin conjugating enzyme. Apc2, Apc11 and Apc10 contribute to a catalytic-substrate binding module linked to a sub-complex of Apc1, Apc4, Apc5 and Cdc23 whose association with Cdc27 is dependent on Cdc16 (ref. 12). Cdc27 engages the IR (Ile-Arg) tails of both Apc10 (ref. 13) and co-activator^{12,14,15}. The small nonessential subunits Apc9, Apc13 (also known as Swm1) and Cdc26 stabilize the TPR sub-complex^{12,16}. Finally, Mnd2 is a budding-yeast-specific subunit implicated in APC/C regulation during meiosis¹⁷.

Structural studies of the APC/C have focused on crystallographic analysis of isolated APC/C subunits and small sub-complexes^{13,18–22}, whereas single particle electron microscopy (EM) has defined the molecular envelope of the APC/C and its complexes with co-activators and the mitotic checkpoint complex^{23–26}. Although approximate locations of the termini of most APC/C subunits have been reported, no systematic assignment of the EM molecular envelope to individual

APC/C subunits has been attempted, limiting our understanding of APC/C molecular mechanisms.

Research on the APC/C has been restricted to the use of native systems. Because most APC/C subunits are essential, genetic manipulations are intrinsically difficult, and the low natural abundance of APC/C has limited structural studies further. Recombinant production of large protein complexes is a significant challenge. Here, we have developed an overexpression system for the APC/C that reconstitutes all 13 APC/C proteins generating a functional E3 ligase. This system allowed us to determine the APC/C mass and subunit stoichiometry accurately and delineate the molecular boundaries of most APC/C subunits within the EM-derived molecular envelope. Integrating crystal structures of TPR subunits^{19–21} and Apc10 (refs 13, 18), and homology models of Apc2 and co-activator, with our cryo-EM reconstruction of an APC/C^{Cdh1-D-box} ternary complex²⁷, provides a high-resolution description of the subunit organization and the framework for generating the first pseudo-atomic model of the APC/C.

Functional assembly of recombinant APC/C

We generated two *Saccharomyces cerevisiae* APC/C sub-complexes—SC8 and TPR5 (Supplementary Fig. 1 and Supplementary Table 2)—the selection of component subunits being guided by the subunit topology of *S. cerevisiae*^{12,16} and *Schizosaccharomyces pombe*²⁷ APC/C. SC8 comprises the core APC/C subunits associated with catalysis and substrate recognition (Apc2, Apc11 and Apc10) together with subunits that are thought to have a structural role (Apc1, Apc4, Apc5 and Cdc23), and Mnd2. We used the MultiBac system^{28,29} to generate a single baculovirus which allowed co-expression and assembly of all eight SC8 subunits in insect cells (Fig. 1a and Supplementary Fig. 1). A second baculovirus, TPR5, contains the TPR subunits Cdc16 and Cdc27, together with the smaller accessory subunits Cdc26, Apc9 and Apc13 (Supplementary Fig. 1). Thus, SC8 and TPR5 together incorporate all 13 APC/C proteins.

¹Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London, SW3 6JB, UK. ²Department of Chemistry, University of Oxford, South Parks Road, Oxford, OX1 3TA, UK.

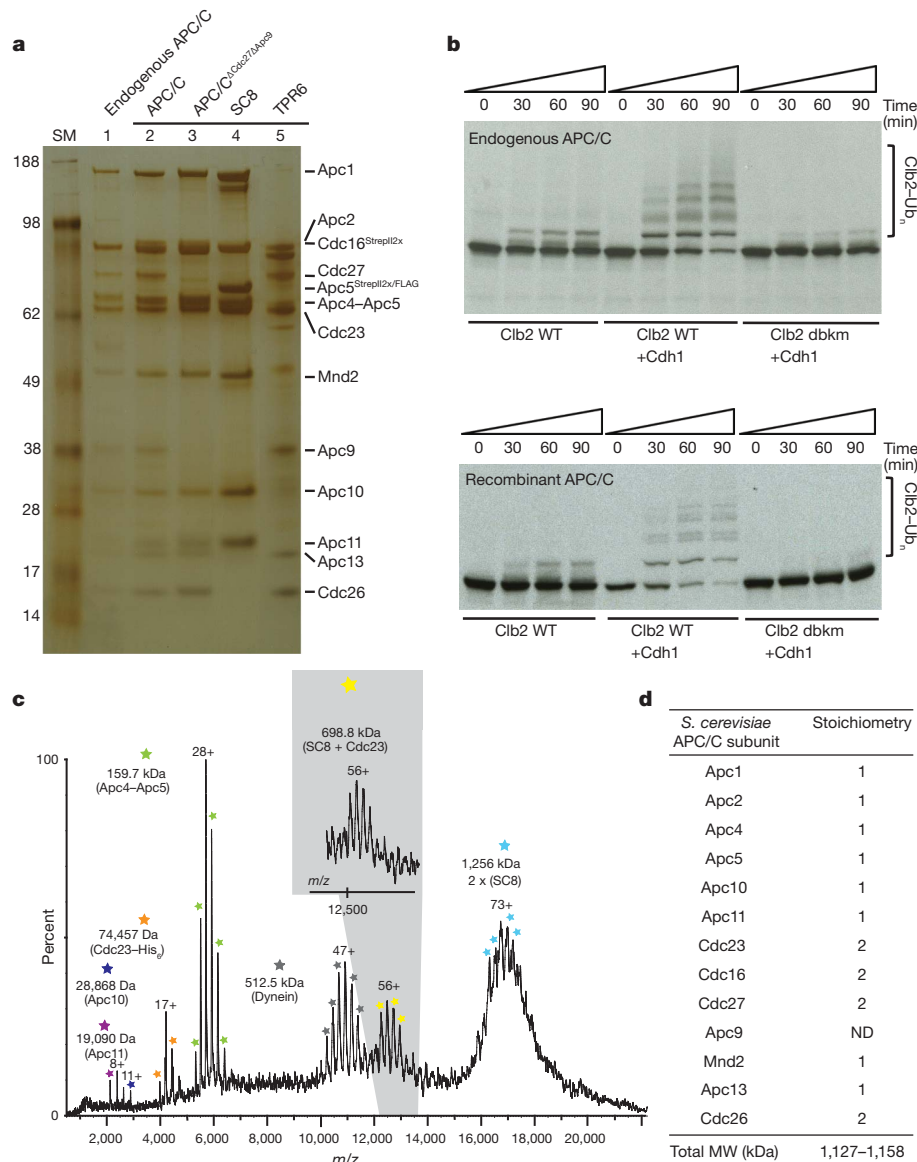


Figure 1 | Recombinant APC/C and sub-complex assembly and purification and nano-electrospray mass spectrometry of SC8. **a**, Silver-stained SDS–PAGE shows comparison of purified endogenous APC/C (lane 1) and purified recombinant *S. cerevisiae* APC/C (lane 2) and APC/C sub-complexes (lanes 3–5). Subunit compositions of sub-complexes are listed in Supplementary Table 2. Endogenous APC/C subunits were identified previously³⁰. **b**, IVT-based ubiquitylation assay showing specific activity of recombinant APC/C (lower panel) depending on the presence of co-activator (*S. cerevisiae* Cdh1) and D-box and KEN-box motifs in the substrate (*S. cerevisiae* Clb2). dbkm,

To generate holo APC/C, SC8 and TPR5 were combined for insect cell co-infection. The resultant recombinant co-expression system yielded ~200-fold more intact APC/C than the endogenous system (0.5 mg l⁻¹ from insect cells compared with 2.5 μg l⁻¹ from budding yeast³⁰). The purified complex was highly homogeneous after a three-step purification procedure (Supplementary Fig. 2). Recombinant APC/C subunits migrate on SDS–polyacrylamide gel electrophoresis (SDS–PAGE) at equivalent positions to their endogenous counterparts (Fig. 1a). Using an *in vitro* transcription/translation-based ubiquitylation assay^{30,31}, we found that recombinant APC/C is active as an E3 ubiquitin ligase towards the mitotic cyclin Clb2, dependent on the presence of co-activator and substrate D-box and KEN-box motifs (Fig. 1b). Thus, the recombinant APC/C recapitulates the activity and specificity of endogenous APC/C³⁰, indicating functional assembly of the complex in the baculovirus–insect cell system.

D-box and KEN-box mutant; WT, wild type. Endogenously purified APC/C was used as a control (upper panel). **c**, Nano-electrospray mass spectrum of SC8. Charge state series corresponding to intact SC8 (yellow), a likely SC8 dimer (light blue), Apc4–Apc5 heterodimer (green) and various individual subunits are indicated. Masses of recombinant proteins are listed in Supplementary Table 1. A dynein contamination is indicated (grey). **d**, Summary of *S. cerevisiae* APC/C subunit stoichiometry as a result of this study. Apo APC/C mass: 1,127 to 1,158 kDa, APC/C^{Cdh1} mass: 1,190 to 1,221 kDa. ND, not determined.

We recorded electron micrographs of recombinant APC/C (Supplementary Fig. 3a). The resultant three-dimensional reconstruction revealed an asymmetric structure with a central cavity defined by an outer lattice-like shell, essentially similar to endogenous APC/C at the current resolution, indicating the correct assembly of recombinant APC/C (Fig. 2a, b and Supplementary Table 3).

APC/C mass and subunit stoichiometry

Accurate information concerning the absolute molecular mass and subunit stoichiometry of the APC/C has been lacking (Supplementary Table 4). To obtain a quantitative estimate of the mass and subunit stoichiometry of the APC/C we applied nano-electrospray mass spectrometry to SC8. Intact SC8 was detected between 12,000 and 13,000 m/z (Fig. 1c). The corresponding charge state series was measured as 698.8 kDa and assigned to an intact SC8, in good agreement

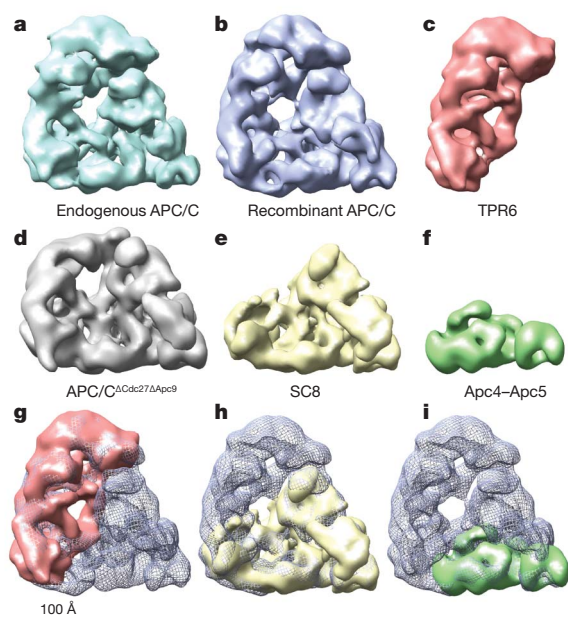


Figure 2 | Three-dimensional EM structure comparisons of recombinant APC/C and APC/C sub-complexes. a–f, EM structure of purified endogenous *S. cerevisiae* holo APC/C²⁷ (a), recombinant holo APC/C (b), TPR6 (c), APC/C^{ΔCdc27ΔApc9} (d), SC8 (e) and Apc4–Apc5 heterodimer (f). g–i, Superimposition of TPR6 (red) (g), SC8 (yellow) (h), and Apc4–Apc5 (green) (i) onto recombinant holo APC/C (blue mesh).

with that predicted for a complex containing all SC8 subunits in unit stoichiometry plus an additional copy of Cdc23 (Supplementary Table 5). Additional charge state series at the lower end of the *m/z* spectrum can be unambiguously assigned to monomeric components of SC8 that have formed in solution, and an Apc4–Apc5 heterodimer (Supplementary Fig. 4), which is a typical by-product of the SC8 purification.

Combining the stoichiometry data for Cdc27 and Cdc16–Cdc26 derived from crystallographic analysis^{19,21}, with mass spectrometry data for SC8 and Apc13 enhanced green fluorescent protein (EGFP) labelling (described below), provides a quantitative estimate of APC/C subunit stoichiometry and molecular mass, with only the absolute stoichiometry of the small budding yeast-specific Apc9 still uncertain. Its association with Cdc27 (ref. 30) raises the possibility that it might exist as two copies per complex. Thus, *S. cerevisiae* APC/C comprises 17 to 18 subunits with a molecular mass in the range of 1,127 to 1,158 kDa (Fig. 1d). These stoichiometries differ from the relative subunit stoichiometries obtained previously using a less accurate radio-iodination method²⁶.

Three-dimensional subunit localization

To assign the three-dimensional molecular boundaries of individual subunits within the APC/C molecular envelope, we generated APC/C sub-complexes lacking defined subunits. We determined their EM structures and compared them with recombinant holo APC/C (Fig. 2). Difference densities can then be assigned to the missing subunits or sub-complexes (Fig. 3).

To map the localization of Cdc27 and Apc9 a complex lacking both subunits was expressed and purified. The resulting APC/C^{ΔCdc27ΔApc9} was identical in composition to holo APC/C except for the missing Cdc27 and Apc9 subunits (Fig. 1a). Fig. 3a shows a superimposition of holo APC/C (Fig. 2b) and APC/C^{ΔCdc27ΔApc9} EM reconstructions (Fig. 2d and Supplementary Fig. 5d). The difference density corresponds to the loss of density in the APC/C^{ΔCdc27ΔApc9} reconstruction relative to holo APC/C and was therefore assigned as Cdc27–Apc9. The density displays an elongated shape with twofold symmetry located at the top of

the APC/C. Mapping Cdc27 to this region agrees with Cdc27 antibody labelling of the head of the TPR lobe^{24,25}. A globular feature at the centre of the Cdc27 density represents the Cdc27 N-terminal dimerization domain, whereas the pair of curved tubular densities represents the two C-terminal TPR super-helices²¹ (Supplementary Table 6).

To determine the full extent of the TPR lobe in the APC/C molecular envelope we generated a larger TPR sub-complex, TPR6 (Fig. 1a), containing Cdc23 in addition to the TPR5 subunits (Supplementary Table 2). EM analysis yielded a quasi twofold symmetrical structure for TPR6 (Fig. 2c and Supplementary Fig. 6). TPR6 adopts an oval bowl-like architecture measuring 180 by 120 by 80 Å, characterized by well defined tubular-like densities organized into a lattice-arrangement (Fig. 2c and Supplementary Fig. 6). We also determined an SC8 EM reconstruction, revealing an asymmetric architecture, also comprising rod-like and globular features (Fig. 2e and Supplementary Figs 7 and 8d). Notably, both TPR6 and SC8 closely match their corresponding regions in the intact APC/C, indicating that these sub-complexes adopt stable autonomous conformations (Fig. 2g, h, Supplementary Figs 6, 8 and Supplementary Table 3).

The difference density obtained by superimposing SC8 (Fig. 2e) onto APC/C^{ΔCdc27ΔApc9} (Fig. 2d) defined the density for Cdc16–Cdc26–Apc13. Fig. 3b shows that the Cdc16–Cdc26–Apc13 difference density bears a striking resemblance to the crystal structure of a Cdc16–Cdc26 heterotetramer²¹. We docked the coordinates of Cdc16–Cdc26 (Fig. 3b) to this density, obtaining near perfect correspondence (Supplementary Table 6). A small difference density feature, not accounted by Cdc16–Cdc26, facing the central TPR cavity, was assigned as Apc13 by EGFP labelling (described below) (Fig. 3b).

SC8 and TPR6 share Cdc23 in common, thus the overlapping density between these two structures after superimposition onto the APC/C reference map can be assigned to Cdc23 (Fig. 3c). The density is dominated by a central globular domain from which two curved tubular features project in opposite directions. We noted a pronounced structural resemblance between the assigned Cdc27 and Cdc23 densities, related by a dyad symmetry operation centred on the Cdc16 dimer axis. Taking Cdc27 as a homology model for Cdc23, we found an almost perfect fit to the assigned Cdc23 density (Fig. 3c and Supplementary Table 6). These findings indicating that Cdc23 is a dimer structurally related to Cdc27 are consistent with our mass spectrometry results revealing two subunits of Cdc23 in SC8 (Fig. 1c, d). Moreover, similar to Cdc27 and Cdc16, multi-angle light scattering showed that the N-terminal region of Cdc23 alone is responsible for its dimerization (Supplementary Fig. 9a).

To locate the small subunit Apc13 we used a C-terminal EGFP tag (Supplementary Fig. 10). In the EM reconstruction we observed a single barrel-like density of EGFP in contact with Apc10 emerging from unassigned density at the Cdc16–Cdc23 interface, corresponding to a small feature in the segmented Cdc16–Cdc26–Apc13 difference density map (Fig. 3b). This indicated that the C- and N-terminal segments of Apc13 interact with the conserved TPR superhelices of Cdc16 and Cdc23, respectively (Supplementary Fig. 11a). A role for Apc13 in mediating the interaction of these two TPR subunits is consistent with Apc13 promoting the stable association of Cdc16 and Cdc27 to the APC/C¹⁶, and biochemical evidence for direct Apc13–Cdc23 interactions³² (Supplementary Fig. 9b).

Assignment of the densities for the three TPR proteins Cdc16, Cdc23 and Cdc27 accounts for approximately 50% of the APC/C molecular mass. Density remaining in SC8 following subtraction of Cdc23 corresponds to Apc1, Apc2, Apc4, Apc5, Apc10, Apc11 and Mnd2. To assign the molecular boundaries of the stable ~155 kDa Apc4–Apc5 heterodimer we determined a three-dimensional EM reconstruction (Fig. 2f and Supplementary Figs 12 and 13). The resultant structure of Apc4–Apc5, revealing an elongated structure with a hook-like feature connected to a ring or disc-like domain, correlates closely with an area within the platform region of APC/C (Figs 2i and 4), consistent with antibody labelling studies of human and fission yeast APC/C^{26,27}.

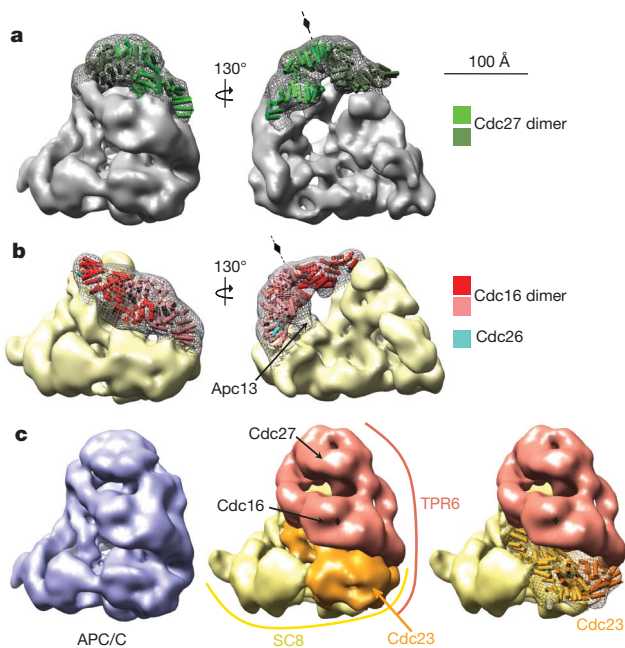


Figure 3 | Three-dimensional localization of TPR subunits and atomic coordinate docking. **a**, Three-dimensional localization of Cdc27–Apc9 by subtracting the APC/C^{ΔCdc27ΔApc9} EM map from the recombinant holo APC/C map. The difference density is drawn as a grey mesh and used as restraints for Cdc27 docking. The N-terminal homo-dimerization domain of *Encephalitozoon cuniculi* Cdc27 (PDB: 3KAE)¹⁹ was docked into the globular density at the top of the APC/C whereas a model of the C-terminal TPR superhelices of Cdc27 were independently docked into the two symmetry related curved tubular densities emerging from the globular density. The two subunits within the homo-dimer are coloured in different shades of green. Molecular envelope shown corresponds to recombinant holo APC/C EM structure. The symmetry axis of the Cdc27 homodimer is indicated in right panel. **b**, Three-dimensional localization of Cdc16–Cdc26–Apc13. The difference density (grey mesh) was calculated by subtracting the SC8 from the APC/C^{ΔCdc27ΔApc9} EM map. The atomic coordinates of the *S. pombe* Cdc16–Cdc26 heterotetramer²¹ were used for rigid body docking. The two Cdc16 subunits within the heterotetramer are shown in red and light red and the Cdc26 N terminus is shown in cyan. Molecular envelope corresponds to the APC/C^{ΔCdc27ΔApc9} EM structure with density assigned to SC8 in yellow surface representation. The symmetry axis of the Cdc16–Cdc26 heterotetramer is indicated in right panel. **c**, Localization of Cdc23. SC8 and TPR6 EM structures were aligned by superimposition onto the recombinant holo APC/C structure (shown in purple, left panel). The resulting overlapping density was coloured in orange (middle panel) and used for the docking of a Cdc23 model derived from the Cdc27 atomic coordinates as modelled in **a**. The two subunits of the homodimeric Cdc23 model are coloured in light and dark orange respectively. The EM densities shown in the middle and right panel are the original SC8 and TPR6 densities aligned to the holo APC/C (left panel). The overlapping Cdc23 density was segmented out from the SC8 EM structure. The symmetry axis of the Cdc16 homodimer is indicated by diamonds.

Combining this definition of Apc4–Apc5 with the previous localization of Apc2 and Apc10 (ref. 27), the remaining unassigned density within the platform region of APC/C can by elimination be assigned to Apc1. In the APC/C EM envelope this density assumes an ‘L’-shape comprising a rod-shaped feature connected to a globular disc-like density that links Apc2 to Cdc23 and incorporates the Apc1 PC repeats (unpublished data; Fig. 4d and Supplementary Fig. 14).

A pseudo-atomic model of the APC/C

Guided by the density assignment of APC/C subunits (Fig. 3) and subunit stoichiometry, we docked APC/C subunit coordinates to the cryo-EM map of the APC/C^{Cdh1–D-box} complex determined at ~11 Å resolution²⁷. The resultant pseudo-atomic model shown in Fig. 4 and Supplementary Movie 1 incorporates Cdc16, Cdc26, Cdc23 and Cdc27 of the TPR sub-complex and, as described previously²⁷, Apc2, Apc11

(N-terminal β-strand) and Apc10, as well as the co-activator Cdh1. Exceptionally good matches between the atomic coordinates and their assigned molecular boundaries independently validates our subunit assignments and the cryo-EM molecular envelope (Supplementary Table 6 and Supplementary Fig. 15). The atomic fitting of the TPR subunits to the cryo-EM map accounts for the major density of the TPR lobe, rationalising its repetitive layered architecture.

Inter-TPR subunit interactions are dominated by the C-terminal TPR superhelices, with few contributions from the dimerization domains, potentially explaining the higher sequence conservation of the C-terminal regions^{19,21} (Supplementary Fig. 11b). These contacts are quite tenuous, comprising a relatively small surface area compared to the TPR dimerization domain interfaces. Unassigned density within the segmented Cdc27–Apc9 density (Figs 3a and 4a) likely corresponds to either Apc9 and/or the variable ~25 kDa inter-TPR motif insertion of *S. cerevisiae* Cdc27 (ref. 19; Fig. 4). Both N- and C-terminal extensions (and the inter-TPR motif insertion of Cdc27) are not represented in existing TPR crystal structures^{19,21}; however, they may also contribute to inter-TPR subunit interactions, and notably, represent the major sites of APC/C phosphorylation^{33,34}. Finally, density protruding from the C-terminal TPR motifs of Cdc23, facing the inner cavity and close to Cdh1 is currently unexplained, but is tentatively assigned to Mnd2 (Supplementary Fig. 16).

Our subunit assignments locate Apc1, Apc4 and Apc5 to the platform of the APC/C (Fig. 4). We assigned Apc5 to the curved hook-like density of the Apc4–Apc5 molecular boundary based on its resemblance to a TPR superhelix (Supplementary Fig. 17). Cdc23 is therefore connected to Apc5 and Apc1, with Apc4 interconnecting Apc1 and Apc5 at the opposite end of the platform. The convex face of TPR motifs 7 to 10 of Cdc23 are responsible for contacting Apc1 and Apc5, explaining how mutations of these motifs disrupt Cdc23 function³⁵ (Supplementary Fig. 11b and 18). Finally, the catalytic substrate-binding module extends, via Apc2 cullin repeats, from Apc1 at the platform to the head of the TPR lobe, positioning Apc10 adjacent to the TPR superhelices of Cdc27.

In the cryo EM map of APC/C^{Cdh1–D-box} (ref. 27) we observe strong density bridging the WD40 domain of Cdh1 and a highly conserved region of the C-terminal domain of Apc2 (Supplementary Fig. 11a,d). This is likely to be the conserved C box of co-activator³⁶, previously proposed to bind the catalytic module through an undefined subunit¹². Our structure therefore implicates this region of Apc2 in C box interactions.

Mechanistic and functional implications

Our pseudo-atomic model of the APC/C allows the integration of genetic, biochemical and structural information. The TPR sub-complex, together with Apc1, Apc4 and Apc5, coordinates the juxtaposition of the catalytic and substrate recognition module subunits Apc2, Apc11 and Apc10 relative to co-activators and APC/C inhibitors. Interestingly, the structurally related C-terminal TPR motifs of Cdc27 and Cdc23 (Supplementary Fig. 18) provide anchor points for co-activator^{14,37} in addition to the putative C box-binding site on Apc2. This APC/C–co-activator interaction network might become partially reorganized during mitosis. For instance, Cdc20 interactions with Cdc27 become essential only after cells progress from prometaphase to metaphase³⁷, thus early in mitosis, the IR tail-binding site on Cdc27 is available for potential interactions with prometaphase substrates. Because the C-terminal Met-Arg sequence of Nek2A, responsible for mediating APC/C interactions³⁸, is structurally related to the IR tails of co-activators and Apc10, the Nek2A MR tail may engage the IR tail-binding site of Cdc27. In contrast, cyclin A is recruited to the APC/C through its binding partner Cks³⁹ that recognizes the phosphorylated TPR lobe.

Unambiguous density corresponding to the RING domain of Apc11 is not visible in the cryo-EM map. One reason for this could be the small size of the RING domain (~7.5 kDa) potentially making

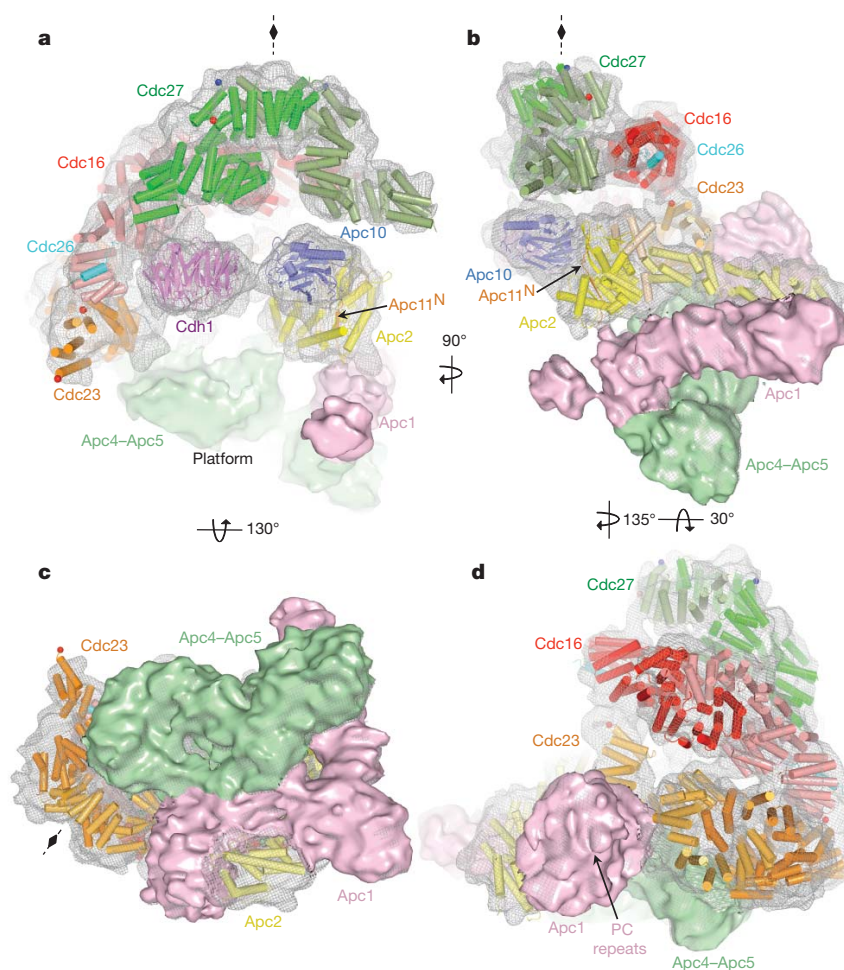


Figure 4 | Subunit organization and pseudo-atomic model of APC/C.

Atomic coordinates of Cdc16–Cdc26, Cdc23, Cdc27, Apc2, Apc10 and Cdh1 were docked in the 11 Å cryo-EM map of the APC/C^{Cdh1-D-box} ternary complex²⁷ represented in grey mesh. A tentative location for Apc11 is indicated by the N-terminal β-strand of Apc11 (labelled Apc11^N), although density for the RING domain is not identified in the cryo-EM map. The surface molecular boundaries of Apc1 (salmon) and Apc4–Apc5 (green) are indicated. Symmetry related monomers of the Cdc16, Cdc23 and Cdc27 homo-dimers are

represented in light and dark red, orange and green, respectively. Local twofold symmetry axes of Cdc27, and Cdc23 are indicated by diamonds. **a**, View onto the central cavity orthogonal to the dyad axis of the Cdc27 homodimer. **b**, **c**, Views related to **(a)** by rotations shown. **d**, View approximately coincident with Cdc16–Cdc26 dyad axis. Red spheres indicate the C termini of Cdc16 and Cdc23, whereas red and blue spheres in Cdc27 denote the N- and C termini of the inter-TPR insert. PC repeats of Apc1 are indicated. Supplementary Movie1 shows an animation of this figure.

it hard to distinguish from other components. An alternative possibility is that in the ternary complex of APC/C, co-activator and substrate, the RING domain is flexible, similar to Rbx1 in activated neddylation Skp1–cullin–F-box⁴⁰. We used our pseudo-atomic APC/C model to interpret the EM structure of the APC/C bound to the mitotic checkpoint complex (MCC)²⁴. MCC (a complex of Cdc20, Bub3, Mad2 and Mad3) interacts mainly with Apc5, and the highly conserved Cdc23 C terminus. In the presence of MCC, Cdc20 is displaced towards Cdc23 away from Cdc27, consistent with biochemical interaction data³⁷. In this displaced position, a direct interaction between the Cdc20 IR tail and C-terminal TPR superhelix of Cdc23 is structurally feasible. Our models for Cdc16 and Cdc23 lack their presumably flexible extreme C termini (Fig. 4). These segments, which become hyper-phosphorylated in early mitosis^{33,34}, are in close proximity to the MCC and proposed co-activator binding site on Cdc23, consistent with the presumed regulatory role of APC/C phosphorylation.

Our recombinant system and the resultant pseudo-atomic model of the APC/C, although not at the accuracy of refined crystal structures, provide insights concerning the molecular mechanisms of this macromolecular machine and a description of how multiple repeat proteins assemble to form an open lattice-like structure. The experimental

approaches described here provide a paradigm for understanding other multi-subunit complexes.

METHODS SUMMARY

Generation of recombinant APC/C and sub-complexes and functional assays. APC/C and sub-complexes were expressed in the baculovirus/insect cell system using the MultiBac system^{28,29}. E3 ligation assays were performed as described^{30,31}.

Electrospray mass spectrometry of intact complexes. Mass spectra were obtained in positive ion mode on a Synapt HDMS instrument (Waters) modified for high-mass operation⁴¹.

Electron microscopy and image analysis. Purified complexes were applied to 2 μm aperture Quantifoil grids coated with continuous thin carbon and negatively stained for electron microscopy at room temperature. Images were recorded in an FEI TF20 electron microscope under low dose conditions using a Tietz F415 CCD camera. Three-dimensional maps were calculated from molecular images using programs from IMAGIC⁴², SPIDER⁴³ and EMAN⁴⁴.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 August; accepted 14 December 2010.

- Peters, J. M. The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nature Rev. Mol. Cell Biol.* **7**, 644–656 (2006).
- Thornton, B. R. & Toczyski, D. P. Precise destruction: an emerging picture of the APC. *Genes Dev.* **20**, 3069–3078 (2006).

3. Sullivan, M. & Morgan, D. O. Finishing mitosis, one step at a time. *Nature Rev. Mol. Cell Biol.* **8**, 894–903 (2007).
4. Glotzer, M., Murray, A. W. & Kirschner, M. W. Cyclin is degraded by the ubiquitin pathway. *Nature* **349**, 132–138 (1991).
5. Pfleger, C. M. & Kirschner, M. W. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.* **14**, 655–665 (2000).
6. Kimata, Y., Baxter, J. E., Fry, A. M. & Yamano, H. A role for the Fizzy/Cdc20 family of proteins in activation of the APC/C distinct from substrate recruitment. *Mol. Cell* **32**, 576–583 (2008).
7. Eytan, E., Moshe, Y., Braunstein, I. & Hershko, A. Roles of the anaphase-promoting complex/cyclosome and of its activator Cdc20 in functional substrate binding. *Proc. Natl Acad. Sci. USA* **103**, 2081–2086 (2006).
8. Passmore, L. A. & Barford, D. Coactivator functions in a stoichiometric complex with anaphase-promoting complex/cyclosome to mediate substrate recognition. *EMBO Rep.* **6**, 873–878 (2005).
9. Sikorski, R. S., Boguski, M. S., Goebel, M. & Hieter, P. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell* **60**, 307–317 (1990).
10. Hirano, T., Kinoshita, N., Morikawa, K. & Yanagida, M. Snap helix with knob and hole: essential repeats in *S. pombe* nuclear protein nuc2⁺. *Cell* **60**, 319–328 (1990).
11. Lupas, A., Baumeister, W. & Hofmann, K. A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). *Trends Biochem. Sci.* **22**, 195–196 (1997).
12. Thornton, B. R. *et al.* An architectural map of the anaphase-promoting complex. *Genes Dev.* **20**, 449–460 (2006).
13. Wendt, K. S. *et al.* Crystal structure of the APC10/DOC1 subunit of the human anaphase-promoting complex. *Nature Struct. Biol.* **8**, 784–788 (2001).
14. Matyskiela, M. E. & Morgan, D. O. Analysis of activator-binding sites on the APC/C supports a cooperative substrate-binding mechanism. *Mol. Cell* **34**, 68–80 (2009).
15. Vodermaier, H. C., Gieffers, C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. TPR subunits of the anaphase-promoting complex mediate binding to the activator protein CDH1. *Curr. Biol.* **13**, 1459–1468 (2003).
16. Schwickart, M. *et al.* Swm1/Apc13 is an evolutionarily conserved subunit of the anaphase-promoting complex stabilizing the association of Cdc16 and Cdc27. *Mol. Cell Biol.* **24**, 3562–3576 (2004).
17. Oelschlaegel, T. *et al.* The yeast APC/C subunit Mnd2 prevents premature sister chromatid separation triggered by the meiosis-specific APC/C-Ama1. *Cell* **120**, 773–788 (2005).
18. Au, S. W., Leng, X., Harper, J. W. & Barford, D. Implications for the ubiquitination reaction of the anaphase-promoting complex from the crystal structure of the Doc1/Apc10 subunit. *J. Mol. Biol.* **316**, 955–968 (2002).
19. Zhang, Z. *et al.* Molecular structure of the N-terminal domain of the APC/C subunit Cdc27 reveals a homo-dimeric tetratricopeptide repeat architecture. *J. Mol. Biol.* **397**, 1316–1328 (2010).
20. Wang, J., Dye, B. T., Rajashankar, K. R., Kurinov, I. & Schulman, B. A. Insights into anaphase promoting complex TPR subdomain assembly from a CDC26-APC6 structure. *Nature Struct. Mol. Biol.* **16**, 987–989 (2009).
21. Zhang, Z., Kulkarni, K., Hanrahan, S. J., Thompson, A. J. & Barford, D. The APC/C subunit Cdc16/Cut9 is a contiguous tetratricopeptide repeat superhelix with a homo-dimer interface similar to Cdc27. *EMBO J.* **29**, 3733–3744 (2010).
22. Han, D. *et al.* Crystal structure of the N-terminal domain of anaphase-promoting complex subunit 7. *J. Biol. Chem.* **284**, 15137–15146 (2009).
23. Dube, P. *et al.* Localization of the coactivator Cdh1 and the cullin subunit Apc2 in a cryo-electron microscopy model of vertebrate APC/C. *Mol. Cell* **20**, 867–879 (2005).
24. Herzog, F. *et al.* Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science* **323**, 1477–1481 (2009).
25. Ohi, M. D. *et al.* Structural organization of the anaphase-promoting complex bound to the mitotic activator Slp1. *Mol. Cell* **28**, 871–885 (2007).
26. Passmore, L. A. *et al.* Structural analysis of the anaphase-promoting complex reveals multiple active sites and insights into polyubiquitylation. *Mol. Cell* **20**, 855–866 (2005).
27. da Fonseca, P. C. *et al.* Structures of APC/C^{Cdh1} with substrates identify Cdh1 and Apc10 as the D-box co-receptor. *Nature* advance online publication, doi:10.1038/nature09625 (24 November 2010).
28. Berger, I., Fitzgerald, D. J. & Richmond, T. J. Baculovirus expression system for heterologous multiprotein complexes. *Nature Biotechnol.* **22**, 1583–1587 (2004).
29. Fitzgerald, D. J. *et al.* Multiprotein expression strategy for structural biology of eukaryotic complexes. *Structure* **15**, 275–279 (2007).
30. Passmore, L. A. *et al.* Doc1 mediates the activity of the anaphase-promoting complex by contributing to substrate recognition. *EMBO J.* **22**, 786–796 (2003).
31. Passmore, L. A., Barford, D. & Harper, J. W. Purification and assay of the budding yeast anaphase-promoting complex. *Methods Enzymol.* **398**, 195–219 (2005).
32. Hall, M. C., Torres, M. P., Schroeder, G. K. & Borchers, C. H. Mnd2 and Swm1 are core subunits of the *Saccharomyces cerevisiae* anaphase-promoting complex. *J. Biol. Chem.* **278**, 16698–16705 (2003).
33. Kraft, C. *et al.* Mitotic regulation of the human anaphase-promoting complex by phosphorylation. *EMBO J.* **22**, 6598–6609 (2003).
34. Steen, J. A. *et al.* Different phosphorylation states of the anaphase promoting complex in response to antimetabolic drugs: a quantitative proteomic analysis. *Proc. Natl Acad. Sci. USA* **105**, 6069–6074 (2008).
35. Sikorski, R. S., Michaud, W. A. & Hieter, P. p62^{Cdc23} of *Saccharomyces cerevisiae*: a nuclear tetratricopeptide repeat protein with two mutable domains. *Mol. Cell Biol.* **13**, 1212–1221 (1993).
36. Schwab, M., Neutzner, M., Mockler, D. & Seufert, W. Yeast Hct1 recognizes the mitotic cyclin Clb2 and other substrates of the ubiquitin ligase APC. *EMBO J.* **20**, 5165–5175 (2001).
37. Izawa, D. & Pines, J. Evidence for how APC-Cdc20 changes its substrate specificity in mitosis. *Nature Cell Biol.* doi:10.1038/ncb2165 (in the press).
38. Hayes, M. J. *et al.* Early mitotic degradation of Nek2A depends on Cdc20-independent interaction with the APC/C. *Nature Cell Biol.* **8**, 607–614 (2006).
39. Wolthuis, R. *et al.* Cdc20 and Cks direct the spindle checkpoint-independent destruction of cyclin A. *Mol. Cell* **30**, 290–302 (2008).
40. Duda, D. M. *et al.* Structural insights into NEDD8 activation of cullin-RING ligases: conformational control of conjugation. *Cell* **134**, 995–1006 (2008).
41. Sobott, F., Hernandez, H., McCammon, M. G., Tito, M. A. & Robinson, C. V. A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal. Chem.* **74**, 1402–1407 (2002).
42. van Heel, M. *et al.* Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.* **33**, 307–369 (2000).
43. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
44. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).

Acknowledgments. This work was funded by a Cancer Research UK grant to D.B. and EU FP7 PROSPECTS (Proteomics Specification in Space and Time) Grant HEALTH-F4-2008-201648 to C.V.R. We thank F. Beuron for help with electron microscopy and Jing Yang for help with insect cell expression.

Author contributions. All authors contributed to experimental design, data analysis and manuscript preparation. A.S. generated recombinant APC/C, APC/C^{ΔCdc27ΔApc9}, SC8 and Apc4–Apc5 complexes and (with R.I.E.) determined their EM structures. Z.Z. generated recombinant TPR6 and Cdc23–Apc13 and performed MALS. P.C.A.F. and E.H.K. collected and analysed TPR6 EM data and P.C.A.F. determined the TPR6 EM structure. A.S. performed ubiquitylation assays. A.S., R.I.E. and P.C.A.F. fitted coordinates. A.S. and F.S. performed mass spectrometry experiments and A.S., F.S. and C.V.R. analysed mass spectrometry data. E.P.M. helped analyse EM data. D.B. directed the project.

Author information. EM maps have been deposited in EMD with accession numbers EMD-1815 (endogenous cryo-EM APC/C^{Cdh1-D-box}), EMD-1816 (endogenous apo APC/C), EMD-1841 (TPR6), EMD-1842 (APC/C^{ΔCdc27ΔApc9}), EMD-1843 (Apc4–Apc5), EMD-1844 (recombinant apo APC/C), and EMD-1845 (SC8). The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.B. (david.barford@icr.ac.uk).

METHODS

Recombinant *S. cerevisiae* APC/C and APC/C sub-complexes cloning using the MultiBac system. The assembly procedure outlined below uses the original MultiBac system as described in²⁸. A summary of the MultiBac transfer vectors generated for holo APC/C co-expression experiments is given in Supplementary Fig. 1.

All 13 full-length *S. cerevisiae* APC/C subunits were PCR-amplified using ultra-high-fidelity Phusion polymerase (Finnzymes) and cloned into pCR4-TOPO vectors (Invitrogen). Purification tags and restriction sites needed for gene assembly into the MultiBac vectors were included in the PCR primers (Sigma). All APC/C gene constructs were sequence verified.

***S. cerevisiae* APC/C gene tagging:** Cdc23 was fused to a non-cleavable C-terminal His₆-tag. Only for Cdc23–Apc13 co-expression experiments a TEV cleavable N-terminal StrepII^{2x}-tag was fused to Cdc23. Apc5 was fused to a C-terminal PreScission cleavable StrepII^{2x}/FLAG-tag⁴⁵ (abbreviated as SF-tag, amino acid sequence: LEVLQ/GPWSHPQFEKGSAGSAGSAGWSHPQFEKGSAGDYKDDDDK) for SC8 expression and purification. Apc5 was cloned untagged for all other co-expression experiments. Cdc16 was fused to a C-terminal PreScission cleavable StrepII^{2x}-tag (LEVLQ/GPWSHPQFEKGSAGSAGSAGWSHPQFEK) for insect cell co-expression experiments with untagged SC8 (abbreviated with SC8^U, untagged refers to Apc5).

In addition to an untagged Apc13 version, Apc13 was fused to a C-terminal EGFP tag by a PCR-based technique called splicing by overlap extension (SOE)^{46,47}.

To remove internal restriction sites in APC/C subunits, which would interfere with the outlined pFBDM and pUCDM assembly procedure (Supplementary Fig. 1), site directed mutagenesis was done by SOE. SpeI restriction sites were deleted in Apc1, Apc5 and Apc2. BstZ17I restriction sites were deleted in Apc5 and Cdc27. An NheI restriction site was removed in Cdc23.

Apc11, Apc4, Mnd2, Apc13, Apc13–EGFP, Apc9 and Cdc27 were cloned via BssHII/NotI for integration into the multiple cloning site (MCS) 1 of the pFBDM plasmid. Apc1, Apc2, Apc5, Apc5^{SF}, Apc10, Cdc16^{StrepII2x}, Cdc26 were cloned via NheI–XmaI for insertion into the MCS2 of pUCDM or pFBDM. Cdc23 was cloned via BssHII–PstI for assembly into pUCDM MCS1.

Stepwise integration of sequence verified APC/C genes into the MultiBac vectors resulted in the following transfer vector constructs: pFBDM–Apc2/Apc11, pFBDM–Apc4/Apc5, pFBDM–Apc4/Apc5^{SF}, pFBDM–Apc10/Mnd2, pFBDM–Cdc16^{StrepII2x}/Apc13, pFBDM–Cdc16^{StrepII2x}/Cdc27, pFBDM–Cdc26, pFBDM–Cdc26/Apc9, pFBDM–Cdc27, pFBDM–Apc13^{EGFP}, pUCDM–StrepII2x/Cdc23/Apc13 and pUCDM–Apc1/Cdc23^{His6}.

Gene cassettes from pFBDM–Apc4/Apc5^{SF}, pFBDM–Apc4/Apc5, pFBDM–Apc10/Mnd2, pFBDM–Cdc26, pFBDM–Cdc26/Apc9, pFBDM–Apc13^{EGFP} and pFBDM–Cdc27 were excised by AvrII–PmeI restriction digest and inserted into the multiplication module (MM) of BstZ17I–SpeI linearized acceptor vectors.

To generate the pFBDM transfer vector for SC8 assembly both Apc4–Apc5 and Apc4–Apc5^{SF} gene cassettes were ligated into the BstZ17I–SpeI-linearized MM of pFBDM–Apc2/Apc11 resulting in pFBDM–Apc2/Apc11/Apc4/Apc5 and pFBDM–Apc2/Apc11/Apc4/Apc5^{SF}, respectively. These resultant vectors were again linearized in the MM and used for the integration of the Apc10–Mnd2 gene cassette.

A similar procedure was applied for pFBDM–TPR5 and pFBDM–TPR5^{Apc13-EGFP} transfer vector assembly: Cdc26/Apc9 was inserted into both BstZ17I–SpeI-linearized pFBDM–Cdc16^{StrepII2x}/Apc13 and pFBDM–Cdc16^{StrepII2x}/Cdc27 resulting in the assembly of pFBDM–Cdc16^{StrepII2x}/Apc13/Cdc26/Apc9 and pFBDM–Cdc16^{StrepII2x}/Cdc27/Cdc26/Apc9. In the final ligation step Apc13–EGFP was inserted into the MM of pFBDM–Cdc16^{StrepII2x}/Cdc27/Cdc26/Apc9 and Cdc27 was integrated into the MM of pFBDM–Cdc16^{StrepII2x}/Apc13/Cdc26/Apc9 resulting in TPR5 and TPR5^{Apc13-EGFP}.

To assemble the pFBDM-based transfer vector for APC/C^{ΔCdc27ΔApc9} expression, the gene cassette containing Cdc26 was ligated into the MM of pFBDM–Cdc16^{StrepII2x}/Apc13. The resulting vector will be referred to as TPR3.

Correct gene assembly into the MultiBac transfer vectors was confirmed by overlapping PCRs and restriction digests.

A summary of the subunit composition of holo APC/C, APC/C^{ΔCdc27ΔApc9}, TPR5, TPR6 and SC8 is given in Supplementary Table 2.

MultiBac bacmid generation. SC8 and SC8^U bacmid generation: the pUCDM–Apc1/Cdc23^{His6} vector was transformed into DH10MultiBac^{Cre} cells (prepared as described in ref. 48). Clonal selection was done as described in ref. 28. The resulting DH10MultiBac cells harbouring a recombinant MultiBac bacmid with an integrated Apc1/Cdc23^{His6} gene cassette were made chemically competent and transformed with pFBDM–Apc2/Apc11/Apc4/Apc5^{SF}/Apc10/Mnd2 or pFBDM–Apc2/Apc11/Apc4/Apc5/Apc10/Mnd2 to generate both tagged and untagged SC8 MultiBac bacmids respectively (SC8 and SC8^U). Cells were incubated with 1 ml

SOC medium at 37 °C for 8 h and plated on L-agar containing chloramphenicol (25 μg ml^{−1}), kanamycin (50 μg ml^{−1}), gentamicin (7 μg ml^{−1}), tetracycline (10 μg ml^{−1}), Bluo-gal (100 μg ml^{−1}) and IPTG (40 μg ml^{−1}). White colonies were selected and restreaked. Bacmid preparation was performed as described⁴⁸.

TPR3 and TPR5 bacmid generation: the pFBDM–Cdc16^{StrepII2x}/Apc13/Cdc26, pFBDM–Cdc16^{StrepII2x}/Apc13/Cdc26/Apc9/Cdc27, and pFBDM–Cdc16^{StrepII2x}/Cdc27/Cdc26/Apc9/Apc13–EGFP were transformed into DH10MultiBac cells to generate TPR3, TPR5 and TPR5^{Apc13-EGFP}. For clonal selection and bacmid preparation see ref. 48.

Cdc23–Apc13 bacmid generation: The pUCDM–StrepII2x/Cdc23/Apc13 vector was transformed into DH10MultiBac^{Cre} cells. For clonal selection and bacmid preparation see ref. 48.

Bacmid verification was done by gene-specific PCRs. Bacmid transfection and virus amplification was done as described in ref. 48.

TPR5^{TEV} and TPR6^{TEV} cloning. The initial TPR5 virus used for TPR5 expression and *in vitro* APC/C reconstitution contained a TEV cleavable StrepII^{2x}-tag at the Cdc16 C terminus (TPR5^{TEV}). The same applies to the TPR6 virus, which contained untagged Cdc23 in addition to the TPR5 subunits (TPR6^{TEV}). APC/C subunits for TPR5^{TEV} and TPR6^{TEV} generation were assembled into the pFBDM transfer vector.

Expression of recombinant APC/C and APC/C sub-complexes. High Five insect cells (Invitrogen) were co-infected with SC8^U and TPR5 (holo APC/C), SC8^U and TPR3 (APC/C^{ΔCdc27ΔApc9}) and SC8^U and TPR5^{Apc13-EGFP} (APC/C^{Apc13-EGFP}) virus at a multiplicity of infection (MOI) of 10 at a cell density 1.6–2.0 × 10⁶ cells per ml. For SC8, TPR5, TPR6 and Cdc23–Apc13 expression High Five insect cells were infected with the corresponding viruses at a MOI of 2 and at a cell density 1.6–2.0 × 10⁶ cells per ml. In all cases High Five cells were incubated at 27 °C at 140 r.p.m. for 72 h.

Purification of recombinant APC/C and APC/C sub-complexes. All purification steps were performed at 4 °C. Cells were harvested at 1,000g for 12 min and frozen in liquid nitrogen. Cell pellets were thawed on ice and resuspended in APC/C lysis buffer (50 mM Tris–HCl pH 8.0, 250 mM NaCl, 5% glycerol, 2 mM DTT, 0.5 mM EDTA, 0.1 mM PMSE, 1 mM benzamide, benzonase (Novagen) and complete EDTA free protease inhibitors (Roche)). After sonication the lysate was spun down for 60 min at 20,000 r.p.m. (Beckman JA-20 rotor), the soluble supernatant was bound to a 5 ml Strep-Tactin Superflow Cartridge (Qiagen) with a flow rate of 2.5 ml min^{−1}. The column was washed with 10 column volumes (CV) of APC/C washing buffer (250 mM NaCl, 50 mM Tris–HCl pH 8.0, 5% glycerol, 2 mM DTT). Recombinant APC/C was eluted with 5 CV APC/C washing buffer supplemented with 2.5 mM desthiobiotin (Sigma). StrepTactin elution fractions were diluted twofold into buffer A (20 mM HEPES–NaOH pH 8.0, 50 mM NaCl, 2% glycerol and 2 mM DTT) and loaded onto a pre-equilibrated ResourceQ anion exchange column (GE Healthcare). The column was washed with 2 CV buffer A and buffer B (20 mM HEPES–NaOH pH 8.0, 1 M NaCl, 2% glycerol and 2 mM DTT) was used for elution. A gradient of 5–40% buffer B was applied over 48 CV. Resource Q peak fractions were loaded onto a Superose 6 10/300 GL column (GE Healthcare) equilibrated in APC/C size exclusion buffer (20 mM HEPES NaOH pH 8.0, 200 mM NaCl, 2 mM DTT, 2% glycerol). In the case of TPR6, the StrepTactin sample was directly loaded onto a TSK-GEL G4000SW (Tosoh Bioscience LLC) size-exclusion column equilibrated in TPR6 size-exclusion buffer (20 mM Tris–HCl pH 7.5, 250 mM NaCl and 1 mM DTT). Cdc23–Apc13 was further purified as described for TPR6, however, loaded onto a Superdex 200 size exclusion column (GE Healthcare).

***S. pombe* Cdc23–Apc13 cloning, expression and purification.** SpCdc23^{1–308} and SpApc13 were PCR-amplified from *S. pombe* cDNA library pTN–TH7 (a gift from T. Nakamura). SpCdc23^(N1–308) was fused to a C-terminal StrepII^{2x}-tag and cloned into the pUCDM transfer vector together with SpApc13. The pUCDM transfer vector were integrated into the MultiBac bacmid by transformation into DH10MultiBac^{Cre} cells. Clonal selection, bacmid purification, insect cell transfection and virus amplification were as described⁴⁸. High Five insect cells were infected with SpCdc23^{1–308}/SpApc13 virus at a MOI of 2 at a cell density between 1.6–2.0 × 10⁶ cells per ml. Protein purification was performed as described for TPR6 apart from the last size exclusion run which was carried out on a Superdex 200 size exclusion column (GE Healthcare).

Endogenous APC/C purification and APC/C ubiquitylation assays. Ubiquitylation assays were performed as described³¹. In this study Clb2 was used as an APC/C substrate. A D-box KEN-box mutant of Clb2 (abbreviated as ‘dbkm’ in Fig. 1b) served as a specificity control.

Multi-angle light scattering (MALS). MALS was performed using a Wyatt MALS system. TPR6 was injected onto a Superose 6 and both SpCdc23–Apc13 and SpCdc23^{1–308} were injected onto an analytical S200 gel filtration column pre-equilibrated in 20 mM Tris–HCl (pH 8.0), 250 mM NaCl, 2 mM DTT, 3 mM NaN₃. Data were analysed using Wyatt Technology software.

Electrospray mass spectrometry of SC8. SC8 was used for MS analysis directly after the first StrepTactin affinity purification step explaining the presence of a single minor contaminant identified as dynein. Prior to MS analysis samples were concentrated using Vivaspin centrifugal concentrators (50 kDa cut-off, Sartorius) to $\sim 2 \mu\text{M}$ and buffer exchanged using Micro Bio-Spin 6 columns (Bio-Rad) into 200 mM ammonium acetate, 0.1% glycerol (pH was adjusted to 8.0 with ammonia solution). Mass spectra were obtained in positive ion mode on a Synapt HDMS instrument (Waters) modified for high-mass operation⁴¹ using a protocol described previously to preserve noncovalent interactions⁴⁹. The following instrumental parameters were applied: nano-electrospray capillary 1,700 V, sample cone 80 V, extractor cone 1 V, backing 5.3 mbar, quadrupole analyser pressure 7.2×10^{-5} mbar and ToF analyser pressure 1.5×10^{-6} mbar. Trap and transfer voltages were kept at 20 and 12 V, respectively. Spectra were externally calibrated using a 33 mg ml^{-1} aqueous solution of caesium iodide (Sigma). Data were acquired and processed with MassLynx software (Waters) and are shown with minimal smoothing. Attempts to acquire mass spectrometry data for apo APC/C and TPR6 were unsuccessful due to protein precipitation and low stability.

Determination of subunit stoichiometry of SC8. The series assigned to SC8 was measured as 698.8 kDa. We estimated the error range as 686.8 to 698.8 kDa, using methods described in ref. 50 which take account of deviations due to residual solvent molecules and buffer ions that may adhere to the complex and potential post-translational modifications. Solvent molecules and buffer ions only contribute positively to the measured mass. We computed all possible subunit compositions within this range based on their theoretical masses listed in Supplementary Table 1 using methods developed in house⁵¹. Assuming that each subunit of the SC8 is present at least once, only six possible solutions fall within this mass range (Supplementary Table 5).

The solution for SC8 comprising all subunits at unit stoichiometry plus one additional copy of Cdc23–His₆ is the most likely assignment due to its consistency with our EM (discussed in text and depicted in Fig. 3c) and MALD data showing that Cdc23 is a homo-dimer (Supplementary Fig. 9). Comparing this calculated mass of 694.1 kDa for the intact (SC8 + Cdc23–His₆) complex with the measured mass of 698.8 kDa we get a calculated deviation of 0.0067 which constitutes an error of 0.67%, well in line with the standards of native mass spectrometry and a value anticipated for a complex of this mass⁵⁰.

Identification of Apc4–Apc5 heterodimer. The predicted mass for an Apc4–Apc5 heterodimer is 159.7 kDa which is in excellent agreement with the charge state series centred between 5,000 and 6,000 m/z (Fig. 1c) measured as 159.7 kDa. Tandem mass spectrometry confirmed the identity of the Apc4–Apc5 charge state series. Collision-induced dissociation resulted in the dissociation of a highly charged Apc4 subunit measuring 75.2 kDa and the corresponding tagged Apc5 subunit (SF-tag) measured as 84.2 kDa (Supplementary Fig. 4).

Electron microscopy and single particle analysis. Electron microscopy of negatively stained APC/C samples. Purified recombinant *S. cerevisiae* APC/C and APC/C sub-complexes were diluted to a concentration of $\sim 50 \mu\text{g ml}^{-1}$ and applied to Quantifoil 2/2 EM grids coated with a second layer of thin carbon. The grids were glow discharged using the Easiglow discharge unit (Pelco) and negatively stained with 2% (w/v) uranyl acetate. The samples were washed twice with APC/C size exclusion buffer before applying the staining solution. Samples were imaged at room temperature in an FEI Tecnai TF20 electron microscope at an accelerating voltage of 200 kV, in low dose mode with an exposure of $\sim 100 \text{ e}^{-} \text{ \AA}^{-2}$, a nominal magnification of $\times 50,000$ and an underfocus chosen to place the first minimum in the contrast transfer function at $\sim 17 \text{ \AA}$. Images were recorded using a Tietz F415 ($4 \text{ k} \times 4 \text{ k}$) CCD camera and fields were binned by a factor of two resulting in a calibrated sampling of 3.47 \AA per pixel.

Image analysis of negatively stained samples. Image processing was performed using IMAGIC⁴², SPIDER⁴³ and EMAN⁴⁴ software. Single molecular images were manually selected using the EMAN boxer program. Examples of EM micrographs and selected particles are shown in Supplementary Figs 3a, 5a, 6a–i, 7a, 10b and 12b. A table of the number of particles picked for each of the different data sets is presented in Supplementary Table 7.

Recombinant holo APC/C EM structure determination. The endogenous APC/C^{Cdh1} map²⁷ was used as the first reference for refinement of the recombinant holo APC/C using a combination of SPIDER (multi-reference alignment) and IMAGIC (classification, projection matching and three-dimensional reconstruction by back-projection) routines. The Cdh1 density in APC/C was used as an internal control to assess reference bias. The refinement procedure encompassed iterative rounds of multi-reference alignment, classification, angular assignment (to selected image-class averages) by projection matching and three-dimensional reconstruction by back-projection. Representative class-averages used in the final three-dimensional reconstruction and their respective re-projections are shown in Supplementary Fig. 3b. After three rounds of refinement the structure converged

to the model shown in Fig. 2b, which is lacking the density for Cdh1 and is clearly similar to the endogenous *S. cerevisiae* APC/C (Fig. 2a).

Recombinant APC/C^{Apc13–EGFP} EM structure determination. The holo APC/C model was used as an initial starting model. Refinement was performed as described for holo APC/C. After the first refinement round a cylindrical density bridging the APC/C inner cavity between the TPR-lobe and Apc10 was already observed. The structure was further refined until the connectivity of this additional density, which accounts for the C-terminal EGFP-tag on Apc13, was stable. To show that this density feature is consistent with the dimensions and shape of a GFP crystal structure (PDB 1GFL) was docked into the assigned density (shown in Supplementary Fig. 10d).

APC/C^{ΔCdc27ΔApc9} EM structure determination. Reference-free EMAN class averages of APC/C^{ΔCdc27ΔApc9} were aligned to APC/C^{Cdh1} re-projections²⁷. Class average three-dimensional re-projection matches showed a consistent density loss at the top of the TPR-lobe, consistent with antibody labelling results for Cdc27 (refs 24, 25). Hence, a small density was removed from the top of the TPR-lobe in the APC/C^{Cdh1} structure. The Cdh1 density was again left intact in the segmented model as a control for reference bias (Supplementary Fig. 5c). Refinement rounds were carried out as described for holo APC/C. After three rounds of refinement the three-dimensional reconstruction no longer changed significantly. The final APC/C^{ΔCdc27ΔApc9} EM structure (Fig. 2d) shows loss of density in the upper part of the TPR-lobe in comparison to the starting model (Supplementary Fig. 5c, d). The Cdh1 density also completely disappeared which further testifies to the lack of reference bias.

Image analysis of negatively stained samples of TPR6 complexes. A preliminary evaluation of the TPR6 data set was carried out by calculating reference-free image class-averages using the refine2d routine from EMAN. Three classes, which were judged to be approximately mutually orthogonal, were selected from the preliminary set for angular assignment using the IMAGIC C1 start-up procedure. These were used to assign angles, by angular reconstitution, to a further selection of 91 classes, which were subsequently back-projected in order to create an *ab initio* three-dimensional map. 75 of these classes were selected, based on their agreement with corresponding re-projections from the *ab initio* map, and used to calculate the map used as the first reference for refinement. The refinement was performed using a combination of IMAGIC and SPIDER software and consisted of multiple rounds of multi-reference alignment, classification, angular assignment (to selected image-class averages) by projection matching and three-dimensional reconstruction by back-projection. The three-dimensional map obtained after six rounds of refinement clearly showed three domains and overall approximate twofold symmetry (Supplementary Fig. 6b–i). Although this map was not fully optimised, it allowed the identification of a corresponding region in the map of native APC/C^{Cdh1}. This region of the APC/C^{Cdh1} map was extracted (Supplementary Fig. 6b–ii) and used as reference for further refinement. In the last round of refinement a total of 1,500 class-averages were calculated, of which 376 were selected to calculate the final three-dimensional map. Examples of image-class averages used in the final reconstruction and their corresponding forward-projections are shown in Supplementary Fig. 6b–iii.

To ensure that the analysis was not biased towards the reference map selected from the APC/C^{Cdh1} map (Supplementary Fig. 6b–ii), two further references were created from the APC/C^{Cdh1} map, one containing extra densities (Supplementary Fig. 6c–i) and the other with densities excluded (Supplementary Fig. 6c–ii). In both cases after just two rounds of refinement the resulting maps converged towards the final map, clearly indicating its reliability (Supplementary Fig. 6b, c).

SC8 EM structure determination. Reference-free EMAN class averages of SC8 were aligned to APC/C^{Cdh1} re-projections. Supplementary Fig. 8b shows the best matching class average re-projection pair. The loss in density in the SC8 class average mapped to a part of the APC/C determined to correspond to TPR6. This information allowed an approximate SC8 segmentation from the APC/C^{Cdh1} map (Supplementary Fig. 8c) that functioned as the starting model for subsequent refinement rounds.

Apc4–Apc5 EM structure determination. The Apc4–Apc5 structure determination was guided by knowledge of the APC/C subunit architecture. The determination of the TPR-lobe architecture (Fig. 3), together with the docking of the catalytic substrate binding module Apc2, Apc11 and Apc10 (ref. 27) significantly restricted the potential localization of Apc4–Apc5 to an APC/C region ('platform region') which links the catalytic sub-complex with the TPR-lobe (Supplementary Fig. 13a).

An initial evaluation of the Apc4–Apc5 data set was carried out by calculating reference-free class-averages using the refine2d routine from EMAN. Despite the small size of the complex ($\sim 155 \text{ kDa}$) high quality EMAN class averages were obtained. A selection of those EMAN class averages were used as references for a SPIDER alignment of the three-dimensional re-projections of the segmented platform region (depicted in yellow surface view in Supplementary Fig. 13a). The best matching class averages three-dimensional re-projection pairs matched

to two different positions within the platform region. Both regions were segmented out and the corresponding densities functioned as starting models (Supplementary Fig. 13b, c) for the first refinement round using 183 manually selected EMAN class averages as input. After four parallel refinement rounds the two models obtained from the two distinct starting models converged to a similar structure, which matched to one side of the platform region comprising a hook like density connected to a disk-shaped feature (Supplementary Fig. 13d, e). Notably, these two models do not clash with the segmented Cdc23 density and the atomic coordinates of the N-terminal Apc2 cullin repeats (Supplementary Fig. 13h).

Docking atomic coordinates into the cryo-EM map of APC/C^{Cdh1-D-box}. Chimera⁵² was used for atomic coordinate docking of APC/C subunits into the segmented negative stain density maps. This docking was further refined using URO software⁵³ (Fig. 3 and Supplementary Table 6). Docking into the cryo-EM map of the APC/C^{Cdh1-D-box} complex (accompanying paper²⁷) was performed using URO software⁵³ (the correlation coefficient obtained for the simultaneous docking of all subunits was 0.778; correlation coefficients for the docking of individual subunits are shown in Supplementary Table 6). The resolution of the cryo-EM map is estimated between 9 and 11 Å by Fourier shell correlation using the 1/2 bit or 0.5 criterion, respectively (Supplementary Fig. 20). The docked coordinates were converted to densities, Fourier low-pass filtered to 9.5 Å and rendered to yield a volume corresponding to their calculated molecular mass, of 506 kDa, assuming a protein density of 0.844 Da Å⁻³. The filtered coordinates were used to guide the rendering of the APC/C^{Cdh1-D-box} map, resulting in a volume corresponding to approximately 1.13 MDa.

Measurement of similarity between the APC/C maps and APC/C subcomplexes. Individual maps were brought onto the same coordinate system using IMAGIC alignment procedures. The same mask was applied to the corresponding region of interest and correlation coefficients were calculated in Chimera.

Variance analysis. A variance map was calculated from the cryo-EM data of the APC/C^{Cdh1-D-box} complex following the bootstrap volume procedure⁵⁴. Analysis showed that the variance stabilized with 100 bootstrap volumes. The variance values were used to colour code the surface of the APC/C^{Cdh1-D-box} map at the chosen contour level using MATLAB. (Supplementary Fig. 19).

Figure preparation. Representations of the EM maps and figures were generated using PyMOL (www.pymol.org) and Chimera⁵² (http://www.cgl.ucsf.edu/chimera). Structural conservation was determined by defining a multiple sequence alignment of APC/C sequences of *S. cerevisiae*, *S. pombe*, *Homo sapiens*, *Drosophila melanogaster* and *Arabidopsis thaliana*. The MSA was used for mapping structural conservation onto crystal and homology-derived coordinates using CONSURF^{55,56}.

45. Gloeckner, C. J., Boldt, K., Schumacher, A., Roepman, R. & Ueffing, M. A novel tandem affinity purification strategy for the efficient isolation and characterisation of native protein complexes. *Proteomics* **7**, 4228–4234 (2007).
46. Warrens, A. N., Jones, M. D. & Lechler, R. I. Splicing by overlap extension by PCR using asymmetric amplification: an improved technique for the generation of hybrid proteins of immunological interest. *Gene* **186**, 29–35 (1997).
47. Heckman, K. L. & Pease, L. R. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nature Protocols* **2**, 924–932 (2007).
48. Fitzgerald, D. J. *et al.* Protein complex expression by using multigene baculoviral vectors. *Nature Methods* **3**, 1021–1032 (2006).
49. Hernández, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nature Protocols* **2**, 715–726 (2007).
50. McKay, A. R., Ruotolo, B. T., Ilag, L. L. & Robinson, C. V. Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. *J. Am. Chem. Soc.* **128**, 11433–11442 (2006).
51. Taverner, T. *et al.* Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. *Acc. Chem. Res.* **41**, 617–627 (2008).
52. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
53. Navaza, J., Lepault, J., Rey, F. A., Alvarez-Rua, C. & Borge, J. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr. D* **58**, 1820–1825 (2002).
54. Penczek, P. A., Yang, C., Frank, J. & Spahn, C. M. Estimation of variance in single-particle reconstruction using the bootstrap technique. *J. Struct. Biol.* **154**, 168–183 (2006).
55. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38** (Suppl. 2), W529–W533 (2010).
56. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33** (Suppl. 2), W299–W302 (2005).

A massive protocluster of galaxies at a redshift of $z \approx 5.3$

Peter L. Capak¹, Dominik Riechers², Nick Z. Scoville², Chris Carilli³, Pierre Cox⁴, Roberto Neri⁴, Brant Robertson², Mara Salvato⁵, Eva Schinnerer⁶, Lin Yan¹, Grant W. Wilson⁷, Min Yun⁷, Francesca Civano⁸, Martin Elvis⁸, Alexander Karim⁶, Bahram Mobasher⁹ & Johannes G. Staguhn¹⁰

Massive clusters of galaxies have been found that date from as early as 3.9 billion years¹ (3.9 Gyr; $z = 1.62$) after the Big Bang, containing stars that formed at even earlier epochs^{2,3}. Cosmological simulations using the current cold dark matter model predict that these systems should descend from ‘protoclusters’—early overdensities of massive galaxies that merge hierarchically to form a cluster^{4,5}. These protocluster regions themselves are built up hierarchically and so are expected to contain extremely massive galaxies that can be observed as luminous quasars and starbursts^{4–6}. Observational evidence for this picture, however, is sparse because high-redshift protoclusters are rare and difficult to observe^{6,7}. Here we report a protocluster region that dates from 1 Gyr ($z = 5.3$) after the Big Bang. This cluster of massive galaxies extends over more than 13 megaparsecs and contains a luminous quasar as well as a system rich in molecular gas⁸. These massive galaxies place a lower limit of more than 4×10^{11} solar masses of dark and luminous matter in this region, consistent with that expected from cosmological simulations for the earliest galaxy clusters^{4,5,7}.

Cosmological simulations predict that the progenitors of present-day galaxy clusters are the largest structures at high redshift^{4,5,7} ($M_{\text{halo}} > 2 \times 10^{11}$ solar masses (M_{\odot}) and $M_{\text{stars}} > 4 \times 10^9 M_{\odot}$ at $z \approx 6$). These protocluster regions should be characterized by local overdensities of massive galaxies on co-moving distance scales of 2–8 Mpc that coherently extend over tens of megaparsecs, forming a structure that will eventually coalesce into a cluster^{4,5,7,9}. Furthermore, owing to the high mass densities and correspondingly high merger rates, extreme phenomena such as starbursts and quasars should preferentially exist in these regions^{4–7,9,10}. Although overdensities have been reported around radio galaxies on ~ 10 –20-Mpc scales^{6,7} and large gas masses around quasars^{11,12} at redshifts greater than $z = 5$, the available data is not comprehensive enough to constrain the mass of these protoclusters and hence provide robust constraints on cosmological models^{6,7,9}.

We used data covering the entire accessible electromagnetic spectrum in the 2-square-degree Cosmological Evolution Survey (COSMOS) field¹³ (right ascension, 10 h 00 min 30 s; declination, $2^{\circ} 30' 00''$) to search for starbursts, quasars and massive galaxies as signposts of potential overdensities at high redshift. This deep, large-area field provides the multiwavelength data required to find protoclusters on scales > 10 Mpc ($5'$). Optically bright objects at redshifts greater than $z = 4$ were identified through optical and near-infrared colours. Extreme star formation activity was found using millimetre-wave^{14,15} and radio¹⁶ measurements, and potential luminous quasars were identified by X-ray measurements¹⁷. Finally, extreme objects and their surrounding galaxies were targeted with the Keck II telescope and the Deep Extragalactic Imaging

Multi-Object Spectrograph (W. M. Keck Observatory, Hawaii) to measure redshifts.

We found a grouping of four major objects at $z = 5.30$ (Fig. 1). The most significant overdensity appears near the extreme starburst galaxy COSMOS AzTEC-3, which contains $> 5.3 \times 10^{10} M_{\odot}$ of molecular gas and has a dynamical mass, including dark matter, of $> 1.4 \times 10^{11} M_{\odot}$ (ref. 8). The far-infrared (60–120- μm) luminosity of this system is estimated to be $(1.7 \pm 0.8) \times 10^{13}$ solar luminosities (L_{\odot}), corresponding to a star formation rate of $> 1,500 M_{\odot}$ per year¹⁸, which is > 100 times the rate of an average galaxy (with luminosity L_{*}) at $z = 5.3$ (ref. 19). The value and error given are the mean estimate and scatter derived from empirical estimates based on the submillimetre flux, radio flux limit, and CO luminosity, along with model fitting. The models predict a much broader range in total infrared (8–1,000- μm) luminosities, ranging from $2.2 \times 10^{13} L_{\odot}$ to $11 \times 10^{13} L_{\odot}$. The large uncertainty results from the many assumptions used in the models, combined with a lack of data constraining the infrared emission at wavelengths less than rest-frame 140 μm . However, the

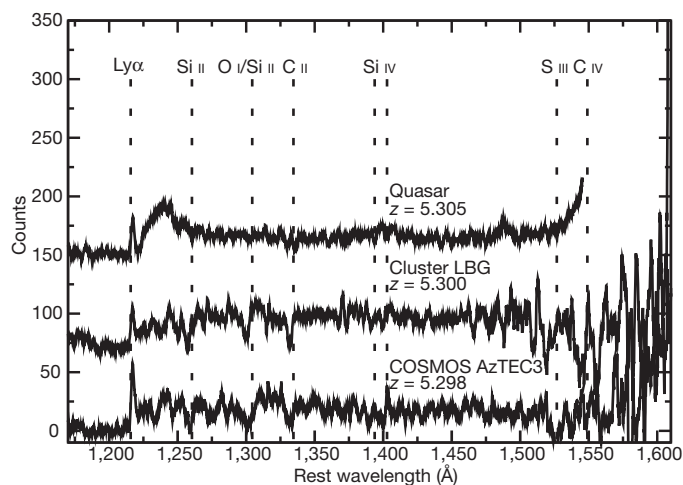


Figure 1 | Spectra of confirmed cluster members. These spectra were taken with the Keck II telescope and correspond to the extreme starburst (COSMOS AzTEC3), a combined spectrum of two Lyman-break galaxies at 95 kpc (Cluster LBG) and the Chandra-detected quasar at 13 Mpc from the extreme starburst. The galaxy spectra show absorption features indicative of interstellar gas (Si II, O I/Si II and C II) and young massive stars (Si IV and C IV) indicative of a stellar population less than 30 Myr old²⁶. The quasar shows broad Lyman- α ($\text{Ly}\alpha$) emission absorbed by strong winds, with a narrow Lyman- α line seen at the same systemic velocity as absorption features in the spectra.

¹Spitzer Science Centre, 314-6 California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. ²Department of Astronomy, 249-17 California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. ³National Radio Astronomy Observatory, PO Box 0, Socorro, New Mexico 87801, USA. ⁴Institut de Radio Astronomie Millimétrique, 300 rue de la Piscine, F-38406 St-Martin-d'Hères, France. ⁵Max-Planck-Institute für Plasma Physics, Boltzmann Strasse 2, Garching 85748, Germany. ⁶Max-Planck-Institute für Astronomie, Königstuhl 17, Heidelberg 69117, Germany. ⁷Department of Astronomy, University of Massachusetts, Lederle Graduate Research Tower B, 619E, 710 North Pleasant Street, Amherst, Massachusetts 01003-9305, USA. ⁸Harvard Smithsonian Center for Astrophysics, 60 Garden Street, MS, 67, Cambridge, Massachusetts 02138, USA. ⁹Department of Physics and Astronomy, University of California, Riverside, California 92521, USA. ¹⁰Johns Hopkins University, Laboratory for Observational Cosmology, Code 665, Building 34, NASA's Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

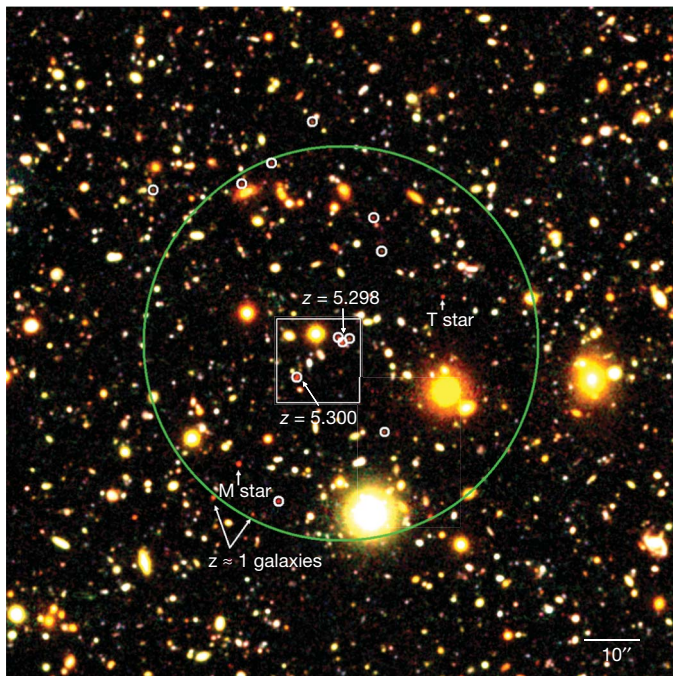


Figure 2 | Image of the region around the protocluster core. This area corresponds to a $2' \times 2'$ region around the starburst (COSMOS AzTEC-3). The $z \approx 5.3$ candidates are marked in white and a 2-Mpc co-moving radius is marked with a green circle. The boxed area is shown larger in Fig. 3, where the optical counterpart of the submillimetre source COSMOS AzTEC-3 is labelled 'Starburst'. Spectroscopic redshifts and other red objects that have been identified as galactic stars or low-redshift galaxies by their spectral energy distribution are also labelled.

observed limit on the submillimetre spectral slope favours models with colder dust and, hence, lower luminosities.

The significance of the overdensity around the starburst is immediately apparent in Figs 2 and 3. In the 1-square-arcmin area ($2.3 \times 2.3 \text{ Mpc}^2$ at $z = 5.3$) around the starburst, we would expect to find 0.75 ± 0.04 bright ($z_{850} < 26$) galaxies with colours consistent with a Lyman break in their spectra at $z = 5.3$ (ref. 19), but instead we find eight. This is an 11-fold overdensity, assuming the redshift range $4.5 < z < 6.5$ probed by typical broadband colour selections^{19,20}.

Within a 2-Mpc radius of the starburst, we find 11 objects brighter than L_* whose intermediate-band colours²¹ are consistent with being at $z = 5.3$. This represents a >11 -fold overdensity in both the measured and the expected density of luminous galaxies. Estimates of the typical variance from clustering and cosmological simulations suggest that this is significant at the $>9\sigma$ level even if we only consider the spectroscopically confirmed systems. Of these 11 objects, three (including the optical counterpart of COSMOS AzTEC-3) are within proper distance of 12.2 kpc ($2''$) of COSMOS AzTEC-3, and two additional spectroscopically confirmed sources are found 95 kpc ($15.5''$) away.

X-ray-selected (0.5–10-keV band) $z > 5$ quasars are extremely rare²² owing to the high luminosities required for detection, yet one is found¹⁷ within 13 Mpc of the starburst at the same spectroscopic redshift as COSMOS AzTEC-3. The distance between these objects is comparable to the co-moving distance scale expected for protoclusters at $z \approx 5$ (refs 5, 7). The optical spectrum of the quasar has deep, blue-shifted gas absorption features indicative of strong winds driven by the energy dissipated from the rapid black-hole growth. The object has an X-ray luminosity of $1.9 \times 10^{11} L_\odot$ and a bolometric luminosity estimated from its spectral energy distribution of $\geq 8.3 \times 10^{11} L_\odot$ (H. Hao *et al.*, manuscript in preparation), implying a black-hole mass of $\geq 3 \times 10^7 M_\odot$ if it is accreting at the Eddington rate, with a more likely mass of $\sim 3 \times 10^8 M_\odot$ for the typical accretion rate of one-tenth the

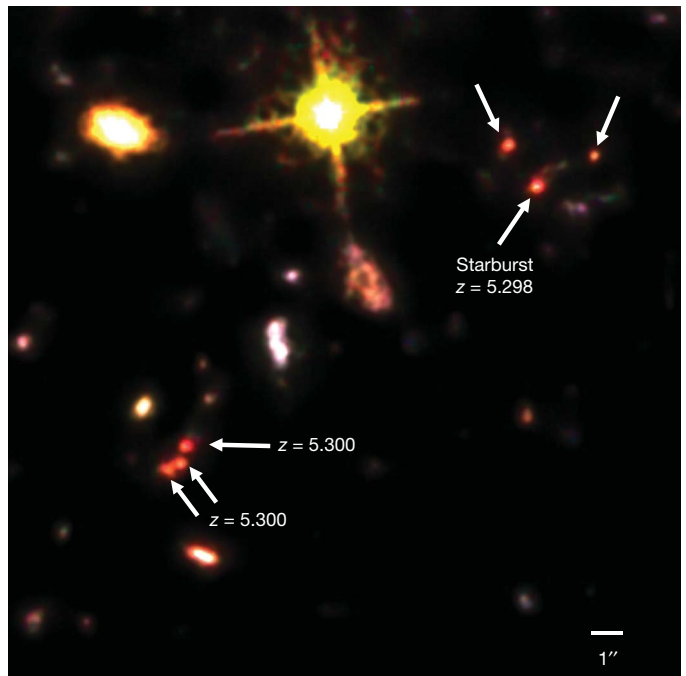


Figure 3 | Detail of the protocluster core. This area corresponds to a $22.5'' \times 22.5''$ area on the sky at a co-moving distance of 0.865 Mpc, or a proper distance of 0.137 Mpc, at $z = 5.298$. The six optically bright objects with spectral energy distributions consistent with $z = 5.298$ are marked and spectroscopic redshifts are indicated. The optical counterpart of the submillimetre source COSMOS AzTEC-3 is labelled 'Starburst'.

Eddington rate⁵. Assuming the final black hole/stellar mass relation to be $M_{\text{BH}} \approx 0.002 M_{\text{stars}}$ (ref. 5) implies that this object will eventually have a stellar mass of $>10^{10} M_\odot$ – $10^{11} M_\odot$, placing it among the most luminous and massive objects at this redshift^{19,23}.

We estimated the stellar mass of the protocluster system by fitting stellar population models to the rest-frame ultraviolet–optical photometry of the individual galaxies in the protocluster. The redshift was fixed at $z = 5.298$, and models with a single recent burst of star formation²⁴ were used, allowing for up to ten visual magnitudes of extinction²⁵. [O II] and H α emission lines were added to the templates with fluxes proportional to the ultraviolet continuum of the template¹⁸. The accuracy of the stellar mass estimate is limited by the sensitivity of the 0.9–2.5- μm photometry. The present data are insufficient to fully break the degeneracy between stellar age and dust obscuration. However, the age of 10 Myr derived from the photometric fitting is consistent with the features seen in the Keck spectra²⁶. Given the range of acceptable fits and the concordance with the Keck spectra, the resulting stellar mass is probably accurate to a factor of ~ 2 (0.3 dex).

Using the described procedure, we conservatively estimate that the starburst AzTEC-3 has a stellar mass of $(1\text{--}2) \times 10^{10} M_\odot$, implying that the baryonic matter is $>70\%$ gas, nearly twice that found in typical starburst systems²⁷ but in agreement with the dynamical estimates^{8,28}. The 11 objects in the protocluster core have a total stellar mass of $>2 \times 10^{10} M_\odot$, with individual galaxies weighing between $0.06 \times 10^9 M_\odot$ and $10 \times 10^9 M_\odot$. With this stellar mass and gas fraction, a lower limit can be placed on the total mass of this system, assuming a global dark matter/baryon ratio of 5.9 (ref. 1). The resulting total halo mass is $>4 \times 10^{11} M_\odot$, with the starburst residing in a halo of mass $>10^{11} M_\odot$, comparable to the halo masses predicted for galaxies that will eventually merge into present-day galaxy clusters⁷. However, we note that the actual mass is probably much higher because much of the baryonic mass is probably in unobserved hydrogen gas, and the starburst object alone accounts for $>37\%$ of the total mass. Furthermore, the contribution of significantly more numerous, fainter (luminosity, $< L_*$) galaxies¹⁹ are not counted in this mass estimate.

The three objects around COSMOS AzTEC-3 probably represent the progenitor of a massive central cluster galaxy (type cD) at lower redshift. These objects are already within the radius of a typical local cD galaxy and their dynamical timescale is ~ 60 Myr, assuming a velocity dispersion of 200 km s^{-1} . Even considering the objects at 95 kpc, the dynamical timescale is less than 0.5 Gyr, providing several dynamical times for a merger to occur by $z \approx 2$ (that is, 2 Gyr later). However, the observed stellar mass in these galaxies is significantly less than the $\sim 10^{11} M_{\odot}$ – $10^{12} M_{\odot}$ in a typical local cD galaxy⁷, indicating that the majority of the stars have yet to form.

The properties of this protocluster are in qualitative and quantitative agreement with galaxy formation simulations^{4,5}. The spatial extent, star formation rate per unit mass and gas properties of the core structure around the extreme starburst are all similar to the predictions for massive-galaxy formation in simulations. Furthermore, the properties of the quasar are also in agreement with the models of the later phases of massive-galaxy formation when the quasar becomes visible. Finally, unlike for previously described overdensities at $z > 5$ (ref. 6), we have strong spectroscopic and photometric evidence for a range of objects including massive, heavily star forming and active galaxies. These are found both in the core of the structure and over a much larger area, indicating that the effects of environment on galaxy formation as early as $z \approx 5$ can be studied. We conclude that this region contains a large-scale baryonic overdensity in the very early Universe that will evolve into a high-mass cluster like those observed at lower redshifts.

Received 18 May; accepted 16 November 2010.

Published online 12 January 2011.

- Larson, D. *et al.* Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: power spectra and WMAP-derived parameters. *Astrophys. J. Suppl. Ser.* (in the press); preprint at (<http://arxiv.org/abs/1001.4635>) (2010).
- Papovich, C. *et al.* A Spitzer-selected galaxy cluster at $z = 1.62$. *Astrophys. J.* **716**, 1503–1513 (2010).
- Mei, S. *et al.* Evolution of the color-magnitude relation in galaxy clusters at $z \sim 1$ from the ACS Intermediate Redshift Cluster Survey. *Astrophys. J.* **690**, 42–68 (2009).
- Springel, V. *et al.* Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435**, 629–636 (2005).
- Li, Y. *et al.* Formation of $z \sim 6$ quasars from hierarchical galaxy mergers. *Astrophys. J.* **665**, 187–208 (2007).
- Overzier, R. *et al.* Stellar masses of Lyman break galaxies, Ly α emitters, and radio galaxies in overdense regions at $z = 4$ – 6 . *Astrophys. J.* **704**, 548–563 (2009).
- Overzier, R. *et al.* Λ CDM predictions for galaxy protoclusters - I. The relation between galaxies, protoclusters and quasars at $z \sim 6$. *Mon. Not. R. Astron. Soc.* **394**, 577–594 (2009).
- Riechers, D. *et al.* A massive molecular gas reservoir in the $z = 5.3$ submillimeter galaxy AzTEC-3. *Astrophys. J. Lett.* **720**, 131–136 (2010).
- Robertson, B. *et al.* Photometric properties of the most massive high-redshift galaxies. *Astrophys. J.* **667**, 60–78 (2007).
- Miley, G. K. *et al.* A large population of ‘Lyman-break’ galaxies in a protocluster at redshift $z \approx 4.1$. *Nature* **427**, 47–50 (2004).
- Walter, F. *et al.* Molecular gas in the host galaxy of a quasar at redshift $z = 6.42$. *Nature* **424**, 406–408 (2003).
- Wang, R. *et al.* Molecular gas in $z \approx 6$ quasar host galaxies. *Astrophys. J.* **714**, 699–712 (2010).
- Scoville, N. Z. *et al.* The Cosmic Evolution Survey (COSMOS): overview. *Astrophys. J. Suppl. Ser.* **172**, 1–8 (2007).
- Scott, K. S. *et al.* AzTEC millimetre survey of the COSMOS field - I. Data reduction and source catalogue. *Mon. Not. R. Astron. Soc.* **385**, 2225–2238 (2008).
- Younger, J. D. *et al.* Evidence for a population of high-redshift submillimeter galaxies from interferometric imaging. *Astrophys. J.* **671**, 1531–1537 (2007).
- Schinnerer, E. *et al.* The VLA-COSMOS Survey. II. Source catalog of the large project. *Astrophys. J. Suppl. Ser.* **172**, 46–69 (2007).
- Elvis, M. *et al.* The Chandra COSMOS Survey. I. Overview and point source catalog. *Astrophys. J.* **184**, 158–171 (2009).
- Kennicutt, J. Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **36**, 189–232 (1998).
- Bouwens, R. J., Illingworth, G. D., Franx, M. & Ford, H. UV luminosity functions at $z \sim 4$, 5, and 6 from the Hubble Ultra Deep Field and other deep Hubble Space Telescope ACS fields: evolution and star formation history. *Astrophys. J.* **670**, 928–958 (2007).
- Hildebrandt, H. *et al.* CARS: the CFHTLS-Archive-Research Survey. II. Weighing dark matter halos of Lyman-break galaxies at $z = 3$ – 5 . *Astron. Astrophys.* **498**, 725–736 (2009).
- Capak, P. *et al.* The first release COSMOS optical and near-IR data and catalog. *Astrophys. J. Suppl. Ser.* **172**, 99–116 (2007).
- Bruza, M. *et al.* High-redshift quasars in the COSMOS survey: the space density of $z > 3$ X-ray selected QSOs. *Astrophys. J.* **693**, 8–22 (2009).
- Stark, D. P. *et al.* The evolutionary history of Lyman break galaxies between redshift 4 and 6: observing successive generations of massive galaxies in formation. *Astrophys. J.* **697**, 1493–1511 (2009).
- Maraston, C. Evolutionary synthesis of stellar populations: a modular tool. *Mon. Not. R. Astron. Soc.* **300**, 872–892 (1998).
- Calzetti, D., Kinney, A. L. & Storchi-Bergmann, T. Dust obscuration in starburst galaxies from near-infrared spectroscopy. *Astrophys. J.* **458**, 132–135 (1996).
- Pettini, M., Steidel, C. C., Adelberger, K. L., Dickinson, M. & Giavalisco, M. The ultraviolet spectrum of MS 1512–CB58: an insight into Lyman-break galaxies. *Astrophys. J.* **528**, 96–107 (2000).
- Tacconi, L. *et al.* High molecular gas fractions in normal massive star-forming galaxies in the young Universe. *Nature* **463**, 781–784 (2010).
- Tacconi, L. J. *et al.* Submillimeter galaxies at $z \sim 2$: evidence for major mergers and constraints on lifetimes, IMF, and CO-H₂ conversion factor. *Astrophys. J.* **680**, 246–262 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements These results are based on observations with: the W. M. Keck Observatory, the IRAM Plateau de Bure Interferometer, the IRAM 30-m telescope with the GISMO 2-mm camera, the Chandra X-ray Observatory, the Subaru Telescope, the Hubble Space Telescope, the Canada-France-Hawaii Telescope with WIRC and MegaPrime, the United Kingdom Infrared Telescope, the Spitzer Space Telescope, the Smithsonian Submillimeter Array Telescope, the James Clerk Maxwell Telescope with the AzTEC 1.1mm camera, and the National Radio Astronomy Observatory's Very Large Array. D.R. and B.R. acknowledge support from NASA through Hubble Fellowship grants awarded by the Space Telescope Science Institute. P.L.C. and N.Z.S. acknowledge grant support from NASA. G.W.W., M.Y. and J.G.S. acknowledge grant support from the NSF.

Author Contributions P.L.C. led the spectroscopic effort, reduced the spectroscopic and photometric data, and led the scientific analysis including the optical and radio/millimetre fitting analysis and cluster properties. N.Z.S. led the spectroscopic and photometric follow-up efforts. D.R., C.C., P.C. and R.N. assisted with the physical interpretation of the radio data. B.R. provided cosmological simulations to check the significance of the protocluster and the likelihood of finding it. M.S., L.Y., M.E., F.C. and B.M. carried out the Keck observations and assisted with the data reduction. E.S. reduced and analysed the radio data. G.W.W. and M.Y. assisted with the submillimetre data analysis. F.C. and M.E. assisted with the X-ray data analysis. A.K. coordinated the 2-mm observations. J.G.S. conducted and reduced the 2-mm observations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.L.C. (capak@astro.caltech.edu).

Observation of scale invariance and universality in two-dimensional Bose gases

Chen-Lung Hung¹, Xibo Zhang¹, Nathan Gemelke^{1†} & Cheng Chin¹

The collective behaviour of a many-body system near a continuous phase transition is insensitive to the details of its microscopic physics; for example, thermodynamic observables follow generalized scaling laws near the phase transition¹. The Berezinskii–Kosterlitz–Thouless (BKT) phase transition^{2,3} in two-dimensional Bose gases presents a particularly interesting case because the marginal dimensionality and intrinsic scaling symmetry⁴ result in a broad fluctuation regime and an extended range of universal scaling behaviour. Studies of the BKT transition in cold atoms have stimulated great interest in recent years^{5–10}, but a clear demonstration of critical behaviour near the phase transition has remained elusive. Here we report *in situ* density and density-fluctuation measurements of two-dimensional Bose gases of caesium at different temperatures and interaction strengths, observing scale-invariant, universal behaviours. The extracted thermodynamic functions confirm the existence of a wide universal region near the BKT phase transition, and provide a sensitive test of the universality predicted by classical-field theory^{11,12} and quantum Monte Carlo calculations¹³. Our experimental results provide evidence for growing density–density correlations in the fluctuation region, and call for further explorations of universal phenomena in classical and quantum critical physics.

In two-dimensional (2D) Bose gases, critical behaviour develops in the BKT transition regime, where an ordered phase with finite-range coherence competes with thermal fluctuations and induces a continuous phase transition from normal gas to superfluid with quasi-long-range order³. In this fluctuation region, a universal and scale-invariant description of the system is expected through the power-law scaling of thermodynamic quantities with respect to the coupling strength and a characteristic length scale^{12,14}—for example, the thermal de Broglie wavelength (Fig. 1a). Especially for weakly interacting gases at finite temperatures, the scale invariance prevails over the normal, fluctuation and superfluid regions because of the density-independent coupling constant¹⁵ and the symmetry of the underlying Hamiltonian⁴.

In this Letter, we verify the scale invariance and universality of interacting 2D Bose gases, and identify BKT critical points. We test the scale invariance of *in situ* density and density fluctuations of 2D gases of ¹³³Cs at various temperatures. We study the universality near the BKT transition by tuning the atomic scattering length using a magnetic Feshbach resonance¹⁶ and observing a universal scaling behaviour of the equation of state and the quasi-condensate density. Finally, by comparing the local density fluctuations and the compressibility derived from the density profiles, we provide strong evidence of a growing density–density correlation in the fluctuation regime.

We begin the experiment by loading a nearly pure ¹³³Cs Bose condensate of $N = 2 \times 10^4$ atoms into a single pancake-like optical potential with strong confinement in the vertical (z) direction and weak confinement in the horizontal (r) direction^{17,18}. The trapping potential, $V(r, z) = m\omega_r^2 r^2/2 + m\omega_z^2 z^2/2$, has mean harmonic trapping frequencies $\omega_r = 2\pi \times 10$ Hz and $\omega_z = 2\pi \times 1,900$ Hz. Here, r denotes the radial distance to the trap centre and m is the caesium atomic mass. In this

trap, the gas reaches temperatures as low as $T = 15$ nK and a moderate peak chemical potential, $\mu_0 < k_B T$. The ratio $\hbar\omega_z/\mu_0 > \hbar\omega_z/k_B T \approx 6$ indicates that the sample is deeply in the 2D regime with $<1\%$ population in the vertical excited states. Here, $\hbar = h/2\pi$, h is the Planck constant, and k_B is the Boltzmann constant. The 2D coupling constant is evaluated according to $g = \sqrt{8\pi}a/l_z$ (ref. 15), where a is the atomic scattering length and $l_z = 200$ nm is the vertical harmonic oscillator length. We control the scattering length a in the range 2–10 nm, resulting in weak coupling strengths $g = 0.05$ – 0.26 . Here, the density-dependent correction to g (refs 15, 19) is expected to be small and negligible ($<2\%$).

We obtain *in situ* density distributions of 2D gases by performing absorption imaging perpendicular to the horizontal plane with a commercial microscope objective and a CCD camera¹⁸ (see Fig. 1b for sample images). About 50 images are collected for each experiment condition, and the average density n and the density variance δn^2 are evaluated pixel-wise (Methods). We obtain the radial density $n(r)$ and variance $\delta n^2(r)$ profiles (Fig. 2 insets) by accounting for the cloud anisotropy and performing azimuthal averaging¹⁷.

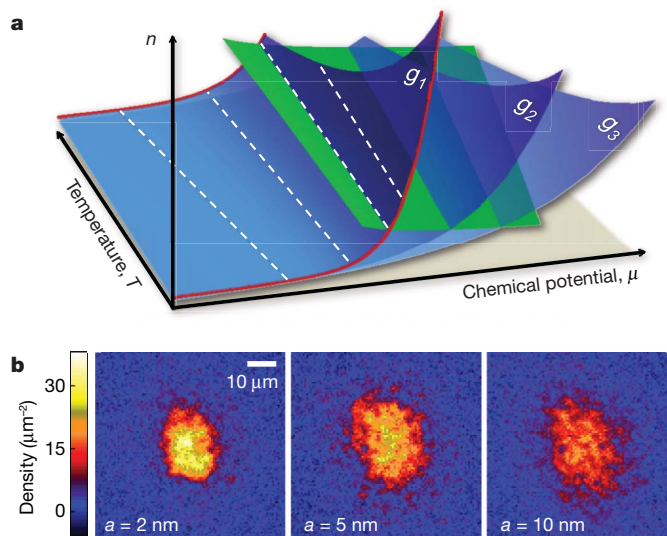


Figure 1 | Illustration of scale invariance and universality in 2D quantum gases. **a**, Scale invariance links any thermodynamic observable at different μ and T via a simple power-law scaling. In a 2D Bose gas with coupling constant $g \ll 1$, atomic density n measured at different temperatures (red lines) can be scaled through constant μ/T and n/T contours (dashed lines). Near the BKT phase transition boundary (green plane), systems with different $g = g_1, g_2, \dots$ (blue planes) scale universally. **b**, *In situ* density measurements of trapped 2D gases provide crucial information to test the hypotheses of scale invariance and universality. Sample images at different scattering lengths a are obtained from single shots.

[†]The James Franck Institute and Department of Physics, University of Chicago, Chicago, Illinois 60637, USA. [†]Present address: Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

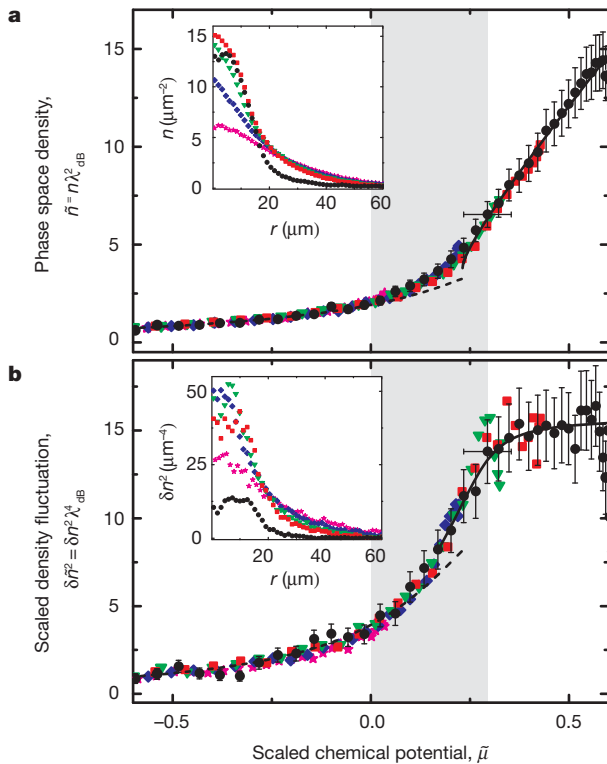


Figure 2 | Scale invariance of density and its fluctuation. **a**, Scaled density (phase space density) $\tilde{n} = n\lambda_{dB}^2$ as a function of the scaled chemical potential $\tilde{\mu} = \mu/k_B T$ measured at five different temperatures: $T = 21$ nK (filled black circles), 37 nK (red squares), 42 nK (green triangles), 49 nK (blue diamonds) and 60 nK (magenta stars), and coupling strength $g = 0.26$. Mean-field expectations for normal gas (dashed line) and superfluid (solid line) are shown for comparison. Inset, radial density profiles before scaling. **b**, Scaled fluctuation $\delta\tilde{n}^2 = \delta n^2 \lambda_{dB}^4$ at different temperatures. Dashed line is the mean-field calculation based on the fluctuation-dissipation theorem²⁵. Solid line is an empirical fit to the crossover feature from which the critical chemical potential $\tilde{\mu}_c$ is determined. Inset, radial fluctuation profiles before scaling. The shaded area marks the fluctuation region $0 < \tilde{\mu} < \tilde{\mu}_c$. Error bars, s.d. of the measurement.

We obtain the equation of state $n(\mu, T)$ from the averaged density profile by assigning a local chemical potential $\mu(r) = \mu_0 - V(r, 0)$ to each point according to the local density approximation. Both T and μ_0 can be determined from the low density wing where the sample is assumed normal and the density profile can be fitted to a mean-field formula $n(\mu, T) = -\lambda_{dB}^{-2} \ln[1 - \exp(\mu/k_B T - gn\lambda_{dB}^2/\pi)]$ (ref. 9), where $\lambda_{dB} = h/\sqrt{2\pi mk_B T}$ is the thermal de Broglie wavelength.

We confirm the scale invariance of a 2D gas by first introducing the dimensionless, scaled forms of density $\tilde{n} = n\lambda_{dB}^2$ (phase space density), fluctuation $\delta\tilde{n}^2 = \delta n^2 \lambda_{dB}^4$, and chemical potential $\tilde{\mu} = \mu/k_B T$, and showing that the equation of state and the fluctuation satisfy the following forms:

$$\tilde{n} = F(\tilde{\mu}) \quad (1)$$

$$\delta\tilde{n}^2 = G(\tilde{\mu}) \quad (2)$$

where F and G are generic functions. This suggests that both energy and length scales are set solely by the thermal energy and the de Broglie wavelength, respectively. An example at $g = 0.26$ ($a = 10$ nm) is shown in Fig. 2. Here we show that while the original density and fluctuation profiles are temperature dependent (Fig. 2 insets), all profiles collapse to a single curve in the scaled units. At negative chemical potential $\tilde{\mu} < 0$, the system is normal and can be described by a mean-field model (dashed lines). In the range $0 < \tilde{\mu} < 0.3$, the system enters the fluctuation regime and deviation from the mean-field calculation becomes evident. Crossing from normal gas to this regime, however, we do not observe a

sharp transition feature in the equation of state. At even higher $\tilde{\mu} > 0.3$, the system becomes a superfluid and the density closely follows a mean-field prediction¹² $\tilde{n} = 2\pi\tilde{\mu}/g + \ln(2\tilde{n}g/\pi - 2\tilde{\mu})$. We notice that the mean-field theory in the superfluid limit cannot also accurately describe the system in the fluctuation regime. Transition into the BKT superfluid phase is most easily seen in the scaled fluctuation $\delta\tilde{n}^2$, which crosses over to a nearly constant value due to the suppression of fluctuation in the superfluid regime²⁰. In the density profile \tilde{n} , a corresponding transition feature can be found when one computes the derivative $\partial\tilde{n}/\partial\tilde{\mu}$, that is, the scaled compressibility $\tilde{\kappa}$, as suggested by the fluctuation-dissipation theorem discussed below. Finally, our measurement suggests that the validity of scale invariance extends to all thermal, fluctuation and superfluid regimes, a special feature of weakly interacting 2D gases⁴ which underlies the analysis of a recent experiment²¹.

We associate the crossover feature in the density fluctuations $\delta\tilde{n}^2$ and the scaled compressibility $\tilde{\kappa}$ with the BKT transition^{20,22}. To estimate the location of the transition point, we apply an empirical fit to this feature and determine the critical chemical potential $\tilde{\mu}_c$ and the critical phase space density \tilde{n}_c (Online Methods). Results at different values of g in the range 0.05 to 0.26 are shown in Fig. 3c, d and compared to the theoretical prediction of $\tilde{n}_c = \ln(\xi/g)$ and $\tilde{\mu}_c = (g/\pi) \ln(\xi_\mu/g)$ (ref. 23), where $\xi = 380$ and $\xi_\mu = 13.2$ are determined from a classical-field Monte Carlo calculation¹¹. Our results show good agreement with the theory, apart from a potential systematic error from the choice of the fit function, which can account for a down shift of 10% in the fitted values of $\tilde{\mu}_c$ and \tilde{n}_c .

Further comparison between profiles at different interaction strengths allows us to test the universality of 2D Bose gases. Sufficiently close to the BKT critical point with $|\tilde{\mu} - \tilde{\mu}_c| < g$, one expects the phase space density to show a universal behaviour¹²:

$$\tilde{n} - \tilde{n}_c = H\left(\frac{\tilde{\mu} - \tilde{\mu}_c}{g}\right) \quad (3)$$

where H is a generic function. Here, density and chemical potential are offset from the critical values \tilde{n}_c and $\tilde{\mu}_c$, which remove the non-universal dependence on the microscopic details of the interaction^{12,14}.

To test the universality hypothesis, we rescale $\tilde{\mu}$ to $\tilde{\mu}/g$ and look for critical values \tilde{n}_c and $\tilde{\mu}_c$ such that the equations of state at all values of g display a universal curve in the phase transition regime (Online Methods). Indeed, we find that all rescaled profiles can collapse to a single curve in the fluctuation region $-1 < (\tilde{\mu} - \tilde{\mu}_c)/g < 0$ and remain overlapped in an extended range of $|\tilde{\mu} - \tilde{\mu}_c|/g \leq 2$ (Fig. 3a), which contrasts with the very different equations of state $\tilde{n}(\tilde{\mu})$ at various g shown in the inset of Fig. 3a. Our result closely follows the classical-field prediction¹² and quantum Monte Carlo calculations¹³ assuming a strictly 2D mean-field contribution; and the fitting parameters (critical density \tilde{n}_c and chemical potential $\tilde{\mu}_c$) show proper dependence on g and are in fair agreement with the prediction of theory¹¹ (Fig. 3c, d). We emphasize that critical values determined from the density fluctuations (Fig. 3c, d) match well with those determined from the universal behaviour, indicating that universality is a powerful tool to determine the critical point from a continuous and smooth density profile. Similar agreement with the theory of critical densities has also been reported on the basis of different experiment techniques^{6,8,10}.

Further universal features near the phase transition can be revealed in the growth of the quasi-condensate density $n_q = \sqrt{n^2 - \delta n^2}$ across the phase transition^{11,12,24}. The quasi-condensate density is a measure of the non-thermal population in a degenerate Bose gas. A finite quasi-condensate density does not necessarily imply superfluidity, but can be responsible for a non-Gaussian distribution observed in momentum space⁸. The quasi-condensate density is predicted to be universal near the critical point, following¹²

$$\tilde{n}_q = Q\left(\frac{\tilde{\mu} - \tilde{\mu}_c}{g}\right) \quad (4)$$

where Q is a generic function and $\tilde{n}_q = n_q \lambda_{dB}^2$.

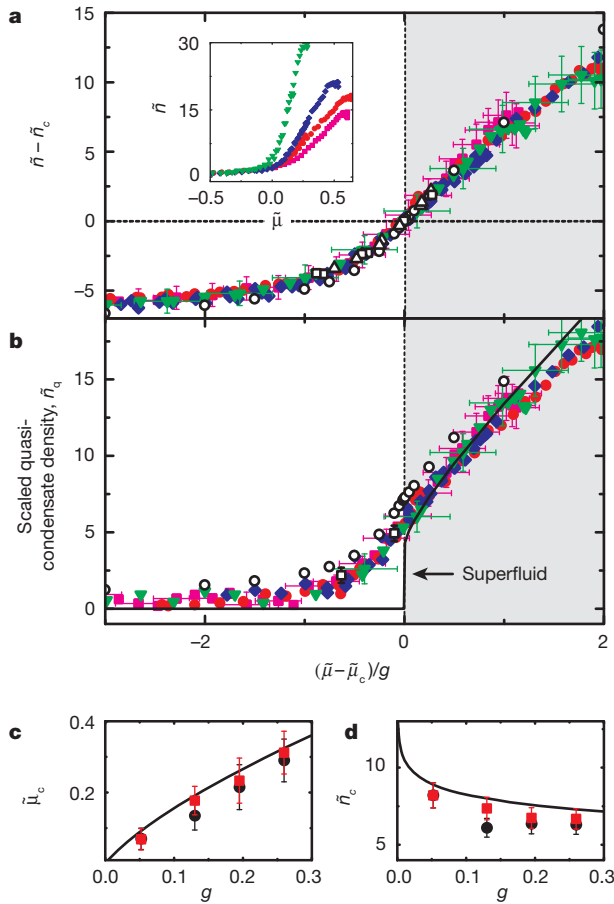


Figure 3 | Universal behaviour near the BKT critical point. **a**, Rescaled density profiles $\tilde{n} - \tilde{n}_c$ measured at various coupling strengths, $g = 0.05$ (filled green triangles), 0.13 (blue diamonds), 0.19 (red circles) and 0.26 (magenta squares). Inset, original equations of state, $\tilde{n}(\tilde{\mu})$. **b**, Scaled quasi-condensate density $\tilde{n}_q = \sqrt{\tilde{n}^2 - \delta\tilde{n}^2}$ at different interaction strengths. In both plots, Monte Carlo calculations from ref. 12 (open circles) and ref. 13 (**a**, open squares for $g = 0.07$ and open triangles for $g = 0.14$; **b**, open squares) are plotted for comparison. The shaded area marks the superfluid regime and the solid line in **b** shows the superfluid phase space density calculation¹². **c**, **d**, Critical values $\tilde{\mu}_c$ and \tilde{n}_c determined from the following methods: universal scaling as shown in **a** (see Online Methods; red squares), density fluctuation crossover (see text; black circles), and Monte Carlo calculations from ref. 11 (solid line). Experiment values coincide at $g = 0.05$ identically, as a result of our analysis (Online Methods). Error bars, s.d. of the measurement.

We use both our density and our fluctuation measurements to evaluate \tilde{n}_q at various g . Adopting $\tilde{\mu}_c$ determined from the universal behaviour of the density profile, we immediately find that all measurements collapse to a single curve in the range $|\tilde{\mu} - \tilde{\mu}_c|/g \leq 2$ with apparent growth of quasi-condensate density entering the fluctuation region (Fig. 3b). The generic function Q we determined is in good agreement with the classical-field¹² and quantum Monte Carlo¹³ calculations with no fitting parameters. Both our density and fluctuation measurements show universal behaviours throughout the fluctuation region where a mean-field description fails, and confirm universality in a 2D Bose gas near the BKT phase transition^{12,13}.

The generic functions we describe above offer new avenues to investigate the critical behaviour of the 2D gas. Following the framework of scale invariance, we compare the dimensionless compressibility $\tilde{\kappa} = \partial\tilde{n}/\partial\tilde{\mu} = F'(\tilde{\mu})$ and the fluctuation $\delta\tilde{n}^2 = G(\tilde{\mu})$ extracted from the measurements at $g = 0.05$ and 0.26 (Fig. 4). In the normal gas regime at low phase space density ($G(\tilde{\mu}), F'(\tilde{\mu}) < 3$), a simple equality $G = F'$ is observed. This result is consistent with the fluctuation-dissipation theorem for a classical grand canonical ensemble²⁵, which gives $k_B T \frac{\partial N}{\partial \mu} = \delta N^2$, where N is the particle number in a detection cell.

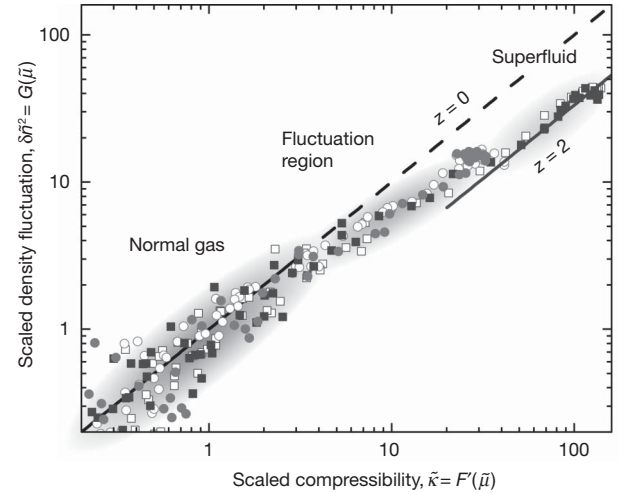


Figure 4 | Fluctuation versus compressibility. Scaled compressibility $\tilde{\kappa} = F'(\tilde{\mu})$ and scaled density fluctuation $\delta\tilde{n}^2 = G(\tilde{\mu})$ are derived from measurements at two interaction strengths, $g = 0.05$ (squares) and $g = 0.26$ (circles), each containing two different temperatures between 20 and 40 nK (filled and open symbols, respectively). Diagonal line shows the expectation of $G = F'$ in the normal gas region. Solid line shows suppressed fluctuation $G = F'/(1+z)$ with $z = 2$. The grey shaded areas mark the normal (left), fluctuation (middle) and superfluid (right) regimes.

In the fluctuation and the superfluid regimes at higher phase space density, our measurement shows that density fluctuations drop below the compressibility, $G < F'$.

Natural explanations for the observed deviation include non-vanishing dynamic density susceptibility at low temperature²⁶ and the emergence of correlations in the fluctuation region²⁷. While the former explanation is outside the scope of this Letter, we show that the correlation alone can explain our observation. Including correlation, the compressibility conforms to^{27,28}:

$$\tilde{\kappa}(\mathbf{r}) = \lambda_{\text{dB}}^{-2} \int \langle \delta\tilde{n}(\mathbf{r}) \delta\tilde{n}(\mathbf{r} + \mathbf{r}') \rangle d^2 r' \quad (5)$$

$$= \delta\tilde{n}^2(\mathbf{r})(1+z) \quad (6)$$

where $\langle \dots \rangle$ denotes ensemble average and $z = \frac{1 + n(\mathbf{r}) \int [g^{(2)}(\mathbf{r}, \mathbf{r} + \mathbf{r}') - 1] d^2 r'}{1 + n(\mathbf{r}) \int [g^{(2)}(\mathbf{r}, \mathbf{r} + \mathbf{r}') - 1] d^2 r'}$ is the relative strength of correlation to local fluctuation $\delta\tilde{n}^2$ (ref. 27). Here $g^{(2)}$ is the normalized second-order correlation function²⁹ and ν denotes the effective area of the resolution limited spot. When the sample is uncorrelated, we have $z = 0$; non-zero z suggests finite correlations in the sample. In the fluctuation region shown in Fig. 4, observing a lower fluctuation than would be indicated by the compressibility, with z approaching 2, suggests that the correlation length approaches or even exceeds our imaging cell dimension, $\sqrt{\nu} \approx 2 \mu\text{m}$. This observation is in agreement with the expected growth of correlation when the system enters the fluctuation region. Similar length scales were also observed in the first-order coherence near the BKT phase transition using an interferometric method⁸ and near the superfluid phase transition in three dimensions³⁰.

In summary, based on *in situ* density measurements at different chemical potential, temperature and scattering length, we have explored and confirmed the global scale invariance of a weakly interacting 2D gas, as well as the universal behaviour near the critical point. Our results provide a detailed description of critical thermodynamics near the BKT transition, and offer new opportunities to investigate other critical phenomena near classical or quantum phase transitions. In particular, we present experimental evidence of the growing correlations in the fluctuation region through the application of the fluctuation-dissipation theorem. Further investigations into the correlations will provide

new insights into the rich critical phenomena near the transition point—for instance, critical opalescence and critical slowing.

METHODS SUMMARY

Preparation and detection of caesium 2D Bose gases are similar to those described in ref. 18. We adjust the temperature of the sample by applying magnetic field pulses near a Feshbach resonance to excite the atoms. We then tune the scattering length to a designated value, followed by 800-ms wait time to ensure full thermalization of the sample.

Absorption imaging is performed *in situ* using a strong resonant laser beam, saturating the sample to reduce the optical thickness. Atom–photon resonant cross-section and atomic density are independently calibrated. Averaged atom number N_i and number fluctuation δN_i^2 at the i th CCD pixel are evaluated pixel-wise based on images taken under identical experiment conditions. The photon shot-noise, weakly depending on the sample's optical thickness, is calibrated and removed from the measured number variance. We correct for the effect of finite imaging resolution on the remaining number variance using calibration from dilute thermal gas measurements. The density fluctuation δn_i^2 is obtained from the recovered atom number variance using $\delta n_i^2 \lambda_{\text{dB}}^2 = \delta N_i^2 / A$, which replaces the dependence on the CCD pixel area A by a proper area scale λ_{dB}^2 (details in Online Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 September; accepted 1 December 2010.

Published online 26 January 2011.

- Stanley, H. E. Scaling, universality, and renormalization: three pillars of modern critical phenomena. *Rev. Mod. Phys.* **71**, S358–S366 (1999).
- Berezinskii, V. L. Destruction of long-range order in one-dimensional and two-dimensional systems possessing a continuous symmetry group — II. Quantum systems. *Sov. Phys. JETP* **34**, 610–616 (1972).
- Kosterlitz, J. M. & Thouless, D. J. Ordering, metastability and phase-transitions in 2 dimensional systems. *J. Phys. Chem.* **6**, 1181–1203 (1973).
- Pitaevskii, L. P. & Rosch, A. Breathing modes and hidden symmetry of trapped atoms in two dimensions. *Phys. Rev. A* **55**, R853–R856 (1997).
- Hadzibabic, Z., Krüger, P., Cheneau, M., Battelier, B. & Dalibard, J. Berezinskii-Kosterlitz-Thouless crossover in a trapped atomic gas. *Nature* **441**, 1118–1121 (2006).
- Krüger, P., Hadzibabic, Z. & Dalibard, J. Critical point of an interacting two-dimensional atomic Bose gas. *Phys. Rev. Lett.* **99**, 040402 (2007).
- Schweikhard, V., Tung, S. & Cornell, E. A. Vortex proliferation in the Berezinskii-Kosterlitz-Thouless regime on a two-dimensional lattice of Bose-Einstein condensates. *Phys. Rev. Lett.* **99**, 030401 (2007).
- Cladé, P., Ryu, C., Ramanathan, A., Helmerson, K. & Phillips, W. D. Observation of a 2D Bose gas: from thermal to quasicondensate to superfluid. *Phys. Rev. Lett.* **102**, 170401 (2009).
- Hadzibabic, Z., Krüger, P., Cheneau, M., Rath, S. P. & Dalibard, J. The trapped two-dimensional Bose gas: from Bose-Einstein condensation to Berezinskii-Kosterlitz-Thouless physics. *N. J. Phys.* **10**, 045006 (2008).
- Tung, S., Lompre, G., Lobser, D., Xia, L. & Cornell, E. A. Observation of the pre-superfluid regime in a two-dimensional Bose gas. *Phys. Rev. Lett.* **105**, 230408 (2010).
- Prokof'ev, N., Ruebenacker, O. & Svistunov, B. Critical point of a weakly interacting two-dimensional Bose gas. *Phys. Rev. Lett.* **87**, 270402 (2001).
- Prokof'ev, N. & Svistunov, B. Two-dimensional weakly interacting Bose gas in the fluctuation region. *Phys. Rev. A* **66**, 043608 (2002).
- Holzmann, M., Chevallier, M. & Krauth, W. Universal correlations and coherence in quasi-two-dimensional trapped Bose gases. *Phys. Rev. A* **81**, 043622 (2010).
- Holzmann, M., Baym, G., Blaizot, J.-P. & Laloë, F. Superfluid transition of homogeneous and trapped two-dimensional Bose gases. *Proc. Natl Acad. Sci. USA* **104**, 1476–1481 (2007).
- Petrov, D. S., Holzmann, M. & Shlyapnikov, G. V. Bose-Einstein condensation in quasi-2D trapped gases. *Phys. Rev. Lett.* **84**, 2551–2555 (2000).
- Chin, C., Grimm, R., Julienne, P. & Tiesinga, E. Feshbach resonances in ultracold gases. *Rev. Mod. Phys.* **82**, 1225–1286 (2010).
- Gemelke, N., Zhang, X., Hung, C.-L. & Chin, C. *In situ* observation of incompressible Mott-insulating domains in ultracold atomic gases. *Nature* **460**, 995–998 (2009).
- Hung, C.-L., Zhang, X., Gemelke, N. & Chin, C. Slow mass transport and statistical evolution of an atomic gas across the superfluid-Mott insulator transition. *Phys. Rev. Lett.* **104**, 160403 (2010).
- Mora, C. & Castin, Y. Ground state energy of the two-dimensional weakly interacting Bose gas: first correction beyond Bogoliubov theory. *Phys. Rev. Lett.* **102**, 180404 (2009).
- Simula, T. P., Davis, M. J. & Blakie, P. B. Superfluidity of an interacting trapped quasi-two-dimensional Bose gas. *Phys. Rev. A* **77**, 023618 (2008).
- Rath, S. P. *et al.* Equilibrium state of a trapped two-dimensional Bose gas. *Phys. Rev. A* **82**, 013609 (2010).
- Holzmann, M. & Krauth, W. Kosterlitz-Thouless transition of the quasi-two-dimensional trapped Bose gas. *Phys. Rev. Lett.* **100**, 190402 (2008).
- Popov, V. N. *Functional Integrals in Quantum Field Theory and Statistical Physics* (Reidel, 1983).
- Bisset, R. N., Davis, M. J., Simula, T. P. & Blakie, P. B. Quasicondensation and coherence in the quasi-two-dimensional trapped Bose gas. *Phys. Rev. A* **79**, 033626 (2009).
- Huang, K. *Statistical Mechanics* 152–154 (Wiley, 1963).
- Kubo, R. The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **29**, 255–283 (1966).
- Toda, M., Kubo, R. & Saitô, N. *Statistical Physics I: Equilibrium Statistical Mechanics* (Springer, 1983).
- Zhou, Q. & Ho, T.-L. Universal thermometry for quantum simulation. Preprint at <http://arXiv.org/abs/0908.3015v2> (2009).
- Naraschewski, M. & Glauber, R. J. Spatial coherence and density correlations of trapped Bose gases. *Phys. Rev. A* **59**, 4595–4607 (1999).
- Donner, T. *et al.* Critical behavior of a trapped interacting Bose gas. *Science* **315**, 1556–1558 (2007).

Acknowledgements We thank Q. Zhou, B. Svistunov, T.-L. Ho, Y. Castin, C.-C. Chien, S. Tung, N. Prokof'ev, J. Freerick and D.-W. Wang for discussions. This work was supported by the NSF (grant numbers PHY-0747907, NSF-MRSEC DMR-0213745), the Packard Foundation, and a grant from the Army Research Office with funding from the DARPA OLE program. N.G. acknowledges support from the Grainger Foundation.

Author Contributions The data were taken and analysed by C.-L.H. The experimental concept was developed by C.-L.H. and C.C. All authors contributed to discussion of the results and preparation of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.-L.H. (clhung@uchicago.edu).

METHODS

Calibration of the atomic surface density and the atom number fluctuation.

Detection of caesium 2D Bose gases is detailed in ref. 18 and the atomic surface density n is evaluated with schemes similar to those discussed elsewhere³¹, where the resonant cross-section σ_0 is independently calibrated using a thin 3D Bose condensate with similar optical thickness and the known atom number-to-Thomas-Fermi radius conversion. The calibrated value of σ_0 can be compared to that determined from the atom shot-noise amplitude in dilute 2D thermal gases, where the noise is evaluated using binned CCD pixels to remove finite resolution effects. For dilute thermal gases, we expect $\delta N^2 = N$, where N is the mean atom number; we compare the fluctuation amplitude to the mean and extract the value of σ_0 . Two results agree to within 10% and the residual nonlinearity in the density calibration is negligible.

We evaluate the atom number variance δN^2 pixel-wise based on images taken under identical experiment conditions. The photon shot-noise contribution δN_p^2 , which weakly depends on the sample's optical thickness $n\sigma_0$, is calibrated and removed from the atom number fluctuation using $\delta N_p^2 = (\delta N_0^2/2)[1 + \frac{(1+\gamma)e^{-n\sigma_0}}{(1+\gamma)^2}]$, where δN_0^2 is the photon shot-noise without atoms and γ is the ratio of the imaging beam intensity to the saturation intensity. Both δN_0^2 and γ are experimentally calibrated. We then correct for the effect of finite resolution on the number fluctuation¹⁷ by comparing the atom number variance in a dilute thermal cloud to its mean atom number, using $\delta N^2 = N$, and applying this calibration to all fluctuations measured at lower temperatures and higher densities.

Density–density correlation in the fluctuation measurement. In the fluctuation measurement, we determine δn^2 from the pixel-wise atom number variance using the formula $\delta n^2 \lambda_{dB}^2 = \delta N^2 / A$, which replaces the dependence on the pixel area A by a natural area scale λ_{dB}^2 . This definition, however, does not fully eliminate the dependence on the imaging resolution spot size $\nu \approx (2 \mu\text{m})^2$. In particular, when the density–density correlation length ξ approaches or exceeds the resolution, the measured fluctuation can depend on the fixed length scale $\sqrt{\nu}$, which can complicate

the scaling behaviour. However, we do not see clear deviation of scale invariance and universality within our measurement uncertainties (Figs 2b and 3b). We attribute this to the small variation of the non-scale-invariant contribution within our limited range of sample temperature. Further analysis of the correlations and fluctuations is in progress and the result will be published elsewhere.

Determination of the BKT critical values from the fluctuation data. We use a hyperbolic function $y(\tilde{\mu}) = s(\tilde{\mu} - \tilde{\mu}_c) - \sqrt{s^2(\tilde{\mu} - \tilde{\mu}_c)^2 + w^2}$ to empirically fit the crossover feature of the density fluctuation near the transition region, assuming $\delta \tilde{n}^2(\tilde{\mu}) = D e^{s(\tilde{\mu})}$, where the critical chemical potential $\tilde{\mu}_c$, the fluctuation in the superfluid regime D , the slope of the exponential rise s , and the width of the transition region w are fitting parameters. The critical phase space density is then determined from the density profile as $\tilde{n}_c = \tilde{n}(\tilde{\mu}_c)$. Other choices of fit functions give similar results, contributing only small systematics from the choice of different models.

Obtaining the universal function. $H(x)$: we use the density profiles in the inset of Fig. 3a to look for critical values \tilde{n}_c and $\tilde{\mu}_c$ such that the equations of state at all values of g collapse to a single universal curve $H(x) = \tilde{n}(\tilde{\mu}) - \tilde{n}_c$, where $x = (\tilde{\mu} - \tilde{\mu}_c)/g$ is the rescaled chemical potential. To do this, we take the profile measured at $g = 0.05 \equiv g_r$ as the reference, evaluate $H_r(x) = \tilde{n}(g_r x + \tilde{\mu}_{c,r}) - \tilde{n}_{c,r}$ using the critical values $\tilde{n}_{c,r}$ and $\tilde{\mu}_{c,r}$ determined from the fluctuation crossover feature, and smoothly interpolate the data to make a continuous reference curve $H_r(x)$ in the range of $|x| \leq 1$. Using this model, we perform minimum χ^2 fits to the profiles measured at all other values of g according to $\tilde{n}(\tilde{\mu}) = \tilde{n}_c + H_r(\frac{\tilde{\mu} - \tilde{\mu}_c}{g})$, with only \tilde{n}_c and $\tilde{\mu}_c$ as free parameters. This procedure successfully collapses all density profiles (see Fig. 3a), and is independent of any theoretical model. The resulting critical values \tilde{n}_c and $\tilde{\mu}_c$ are plotted in Fig. 3c, d.

- Reinaudi, G., Lahaye, T., Wang, Z. & Guéry-Odelin, D. Strong saturation absorption imaging of dense clouds of ultracold atoms. *Opt. Lett.* **32**, 3143–3145 (2007).

Programmable nanowire circuits for nanoprocessors

Hao Yan^{1*}, Hwan Sung Choe^{2*}, SungWoo Nam^{3*}, Yongjie Hu¹, Shamik Das⁴, James F. Klemic⁴, James C. Ellenbogen⁴ & Charles M. Lieber^{1,3}

A nanoprocessor constructed from intrinsically nanometre-scale building blocks is an essential component for controlling memory, nanosensors and other functions proposed for nanosystems assembled from the bottom up^{1–3}. Important steps towards this goal over the past fifteen years include the realization of simple logic gates with individually assembled semiconductor nanowires and carbon nanotubes^{1,4–8}, but with only 16 devices or fewer and a single function for each circuit. Recently, logic circuits also have been demonstrated that use two or three elements of a one-dimensional memristor array⁹, although such passive devices without gain are difficult to cascade. These circuits fall short of the requirements for a scalable, multifunctional nanoprocessor^{10,11} owing to challenges in materials, assembly and architecture on the nanoscale. Here we describe the design, fabrication and use of programmable and scalable logic tiles for nanoprocessors that surmount these hurdles. The tiles were built from programmable, non-volatile nanowire transistor arrays. Ge/Si core/shell nanowires¹² coupled to designed dielectric shells yielded single-nanowire, non-volatile field-effect transistors (FETs) with uniform, programmable threshold voltages and the capability to drive cascaded elements. We developed an architecture to integrate the programmable nanowire FETs and define a logic tile consisting of two interconnected arrays with 496 functional configurable FET nodes in an area of $\sim 960 \mu\text{m}^2$. The logic tile was programmed and operated first as a full adder with a maximal voltage gain of ten and input–output voltage matching. Then we showed that the same logic tile can be reprogrammed and used to demonstrate full-subtractor, multiplexer, demultiplexer and clocked D-latch functions. These results represent a significant advance in the complexity and functionality of nanoelectronic circuits built from the bottom up with a tiled architecture that could be cascaded to realize fully integrated nanoprocessors with computing, memory and addressing capabilities.

The programmable nanowire FETs (NWFETs) incorporated a top-gated geometry (Fig. 1a, left panel) using Ge/Si core/shell nanowires as the semiconductor channel because previous work¹² had shown that this provided high yields of devices with uniform threshold voltages and on-current characteristics. To realize programmable, non-volatile NWFETs, we implemented a trilayer Al_2O_3 – ZrO_2 – Al_2O_3 dielectric structure (Fig. 1a, right-hand panels) for charge trapping¹³. For a p-type Ge/Si nanowire channel, negative trapped charges increase the hole density (Fig. 1a, top right) and positive trapped charges decrease the hole density (Fig. 1a, bottom right) in the channel. The modulation of carrier density by trapped charges shifts the threshold of the NWFET in a predictable and non-volatile manner. We grew the Al_2O_3 – ZrO_2 – Al_2O_3 dielectric structure by atomic-layer deposition after fabrication of metallic source and drain nanowire contacts (Methods). A cross-sectional transmission electron microscopy image recorded from a representative device (Fig. 1b) shows that our NWFET device consists of the designed structure with a 10-nm-diameter germanium nanowire core, a 2-nm-thick concentric silicon shell and conformal 2-nm Al_2O_3 , 5-nm ZrO_2 and 5-nm Al_2O_3 layers.

The gate response of a NWFET with a trilayer dielectric was characterized in a device with six gate lines, a 1×6 node element (Fig. 1c, inset). For these measurements, we used one gate line as the active gate and the other gate lines were grounded. The drain–source current, I_{ds} , recorded as a function of drain–source voltage, V_{ds} , for different values of gate voltage, V_{gs} (Fig. 1c), has the behaviour expected of a p-type depletion-mode FET¹⁴. The conductance– V_{gs} curves of the same device with ± 6 -V (Fig. 1d, blue) and ± 9 -V (Fig. 1d, red) sweeps in V_{gs} show anticlockwise hysteresis loops that agree well with the charge-trapping mechanism¹⁵. The hysteresis window increases by ~ 2 V in the ± 6 to ± 9 -V V_{gs} sweeps, which is consistent with more charge being trapped at larger voltages and the charge-trapping model¹³. Significantly, these data demonstrate that two distinct states are observed. After a gate bias of -6 V, the conductance of the NWFET changed by $>10^3$ as V_{gs} varied between 0 and 2 V; in contrast, after a

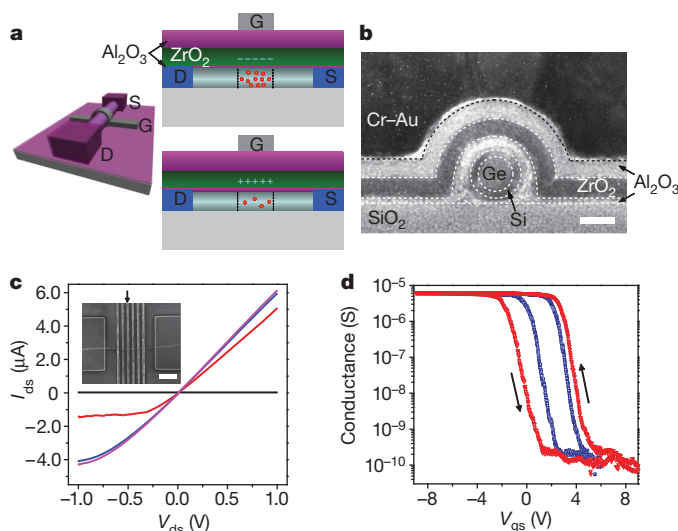


Figure 1 | Structure and characterization of the programmable NWFET. **a**, Left: schematic of the top-gated NWFET; S, D and G correspond to source, drain and gate, respectively. Right: representative hole concentration in a p-type Ge/Si NWFET for two charge-trapping states illustrating carrier accumulation for a negative trapped charge (top right) and depletion for a positive trapped charge (bottom right) in the ZrO_2 layer. **b**, Cross-sectional transmission electron microscopy image of a representative nanowire device, with substrate surface (SiO_2) and gate (Cr–Au) at the bottom and the top of the image, respectively. Other components of the nanowire and dielectric layers are labelled, and dashed lines define the boundary between different components. Scale bar, 10 nm. **c**, I_{ds} – V_{ds} curves recorded from a six-gate NWFET with $V_{\text{gs}} = 8$ (black), 3 (red), 0 (blue) and -8 V (magenta) (G3), and G1, G2, and G4–G6 grounded. Inset, scanning electron microscopy image of the device. The small black arrow indicates G3. Scale bar, 1 μm . **d**, Semi-logarithmic plot of conductance versus V_{gs} for the same device as in **c**, recorded for ± 6 -V (blue) and ± 9 -V (red) sweeps at $V_{\text{ds}} = 0.5$ V; arrows represent sweep/hysteresis direction.

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ²Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. ³School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. ⁴Nanosystems Group, The MITRE Corporation, McLean, Virginia 22102, USA.

*These authors contributed equally to this work.

gate bias of +6 V, the conductance change was less than 50% over the same V_{gs} range. We thus define the former state as 'active', because the NWFET behaves like an active transistor, and define the latter state as 'inactive', because the device behaves like a passive interconnection. Neither programmed state shows degradation on the timescale of a day (Supplementary Fig. 1). The stable programmability of individual NWFETs between the active and inactive states allows distinct functional circuits to be realized from arrays as described below.

We initially investigated the potential of these multi-input programmable NWFETs for building integrated circuits with two coupled nanowire elements (Fig. 2a, left panel), where the first element, NW1, had four independently configurable input gates, G1–G4, and the second element, NW2, had a single input gate connected to the output (drain) of NW1. In this demonstration, the first and third gate nodes of NW1 and the gate node of NW2 were set to the active state (Fig. 2a, green dots), and the other gate nodes were set to the inactive state (Methods). With source voltages of 2.5 and 3 V applied to NW1 and NW2, respectively, input G1 was switched between 0 and 1 V while G2–G4 were held at 0 V (Fig. 2a,

top right). Notably, simultaneous measurements of the output voltage from NW1, V_{IG} , and NW2, V_{out} , with the 0- and 1-V G1 input variations (Fig. 2a, lower right), show that V_{IG} is switched between high (2.2-V) and low (0.2-V) levels and that V_{out} is toggled between low (0.6-V) and high (3.0-V) levels. Similar switching of the V_{IG} and V_{out} levels was recorded when the input to the other active node, G3, was varied and G1, G2, G4 held at 0 V. However, no switching of the V_{IG} and V_{out} levels was observed when the input voltage on either of the inactive input nodes, G2 and G4, was changed from 0 to 1 V. These results show that our programmable NWFET functions as a transistor switch in its active state and that multiple switches can be coupled together by feeding the output of one FET into the input gate of another, and thus suggest that assembly of programmable NWFETs into a suitable architecture could yield integrated circuits capable of processing.

To exploit the unique properties of our programmable NWFETs, while simultaneously recognizing assembly limitations, we have developed a scalable system architecture in which both the locations and the interconnections of transistors are decided after fabrication. This architecture was formulated with the concept of building extended nanoprocessor systems consisting of arrays of interconnected logic tiles^{10,16} (Supplementary Fig. 2). The unit logic tile (Fig. 2b), refined both by extensive simulation¹¹ and by experiment, consists of two programmable, non-volatile nanowire transistor arrays (PNNTAs). The tile is sized to be able to execute a program equivalent to a small number of logic gates, and functions as follows. Metal electrodes are used to gate nanowires in the block-1 PNNTA (Fig. 2b, upper left), and the output of the nanowires is connected by metal electrodes to static load devices. By programming selected nanowire gate nodes to the active transistor state, NOR logic gates⁵ can be mapped into block 1. The outputs of this NOR logic circuit are passed over and used as gate inputs to the block-2 PNNTA (Fig. 2b, lower right) that is also programmed with NOR logic gates. In this way, the outputs of the logic circuits in block 1 can be used to drive the circuit in block 2, thus making it possible to form two-level networks of logic gates in the unit tile that represent arbitrary Boolean functions.

We realized the key architectural tile by fabricating PNNTAs as shown schematically in Fig. 3a (Methods). Briefly, a parallel array of Ge/Si nanowires was assembled by shear-printing¹⁵, source and drain electrodes were defined by electron beam lithography (EBL), atomic-layer deposition was used to deposit the Al_2O_3 - ZrO_2 - Al_2O_3 charge-trapping structure and then a second step of EBL was used to define input gate lines. In this way, two blocks of NWFETs were fabricated, namely block 1 (Fig. 3a, left) and block 2 (Fig. 3a, right), which correspond to the unit tile of our PNNTA architecture (Fig. 2b). Dark-field optical microscopy and scanning electron microscopy images (Supplementary Fig. 3) reveal a total of 496 programmable NWFET devices laid out in two separate arrays with a total area of $\sim 960 \mu\text{m}^2$, where each device node consists of a single nanowire crossed by a gate line. The average area per node, $\sim 1.9 \mu\text{m}^2$, is relatively large in these proof-of-concept studies but does not represent a lower limit, as previous studies demonstrating close-packed nanowire assembly¹⁷ and the scaling of charge-trapping devices¹⁸ indicate that an area 10^3 -fold smaller, $\sim 0.0017 \mu\text{m}^2$, is achievable.

To realize functional logic with the PNNTA tile requires uniform device characteristics among individual nanowire elements. Specifically, the deviation of the threshold voltage, V_{th} , in both the active and the inactive state must be smaller than the difference in V_{th} between the two states. We characterized the V_{th} values of 70 NWFET nodes from block 1 of the fabricated PNNTA structure in both the active and the inactive state (Fig. 3b). Notably, we found that 60 of 70 nodes (86%) in the active state had V_{th} values ≤ 2 V and that 61 of 70 nodes (87%) in the inactive state had V_{th} values ≥ 3.5 V (Fig. 3b). The high yield of NWFET devices reflects the uniformity of the Ge/Si nanowire building block¹², and controlled assembly¹⁵ allows any defective elements to be excluded readily from the functional circuit. For the demonstration of logic

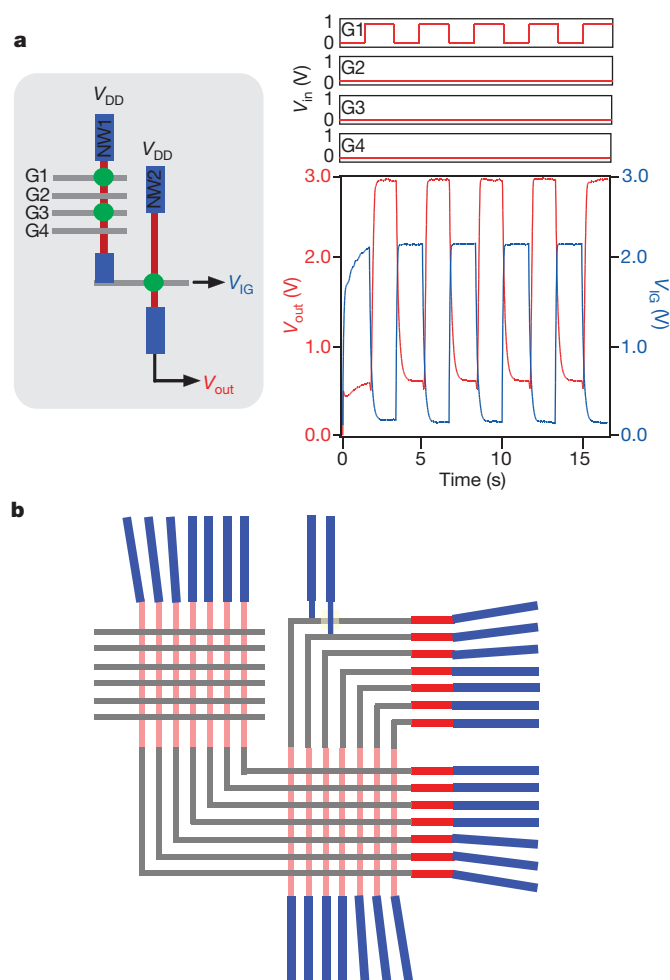


Figure 2 | Coupled NWFET devices and PNNTA architecture.

a, Characterization of a nanowire-nanowire, coupled multigate device. Left: schematic of the device. Green dots indicate the gate nodes that were programmed as an active state. Top right: input signals to G1–G4. Bottom right: output signals from NW1 (V_{IG} , blue) and NW2 (V_{out} , red). **b**, Design of the unit logic tile for integrated nanoprocessors containing two PNNTAs, block 1 (upper left) and block 2 (lower right), comprising charge-trapping nanowires (pink) and metal gate electrodes (grey). The PNNTAs are connected to two sets of load devices (red). Lithographic-scale electrodes (blue) are integrated for input and output. Each PNNTA provides programmable logic functionality of up to approximately eight distinct logic gates. More-complex logic functions can be computed through the hierarchical interconnection of unit logic tiles in linear arrays (Supplementary Fig. 2).

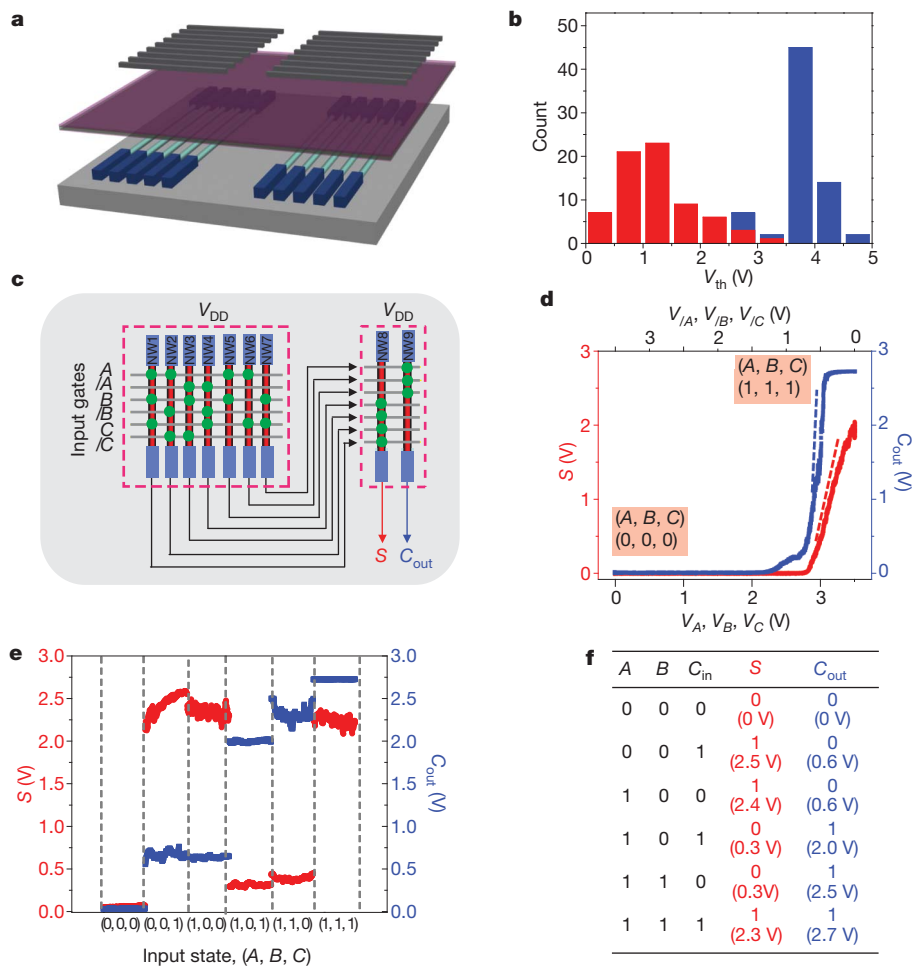


Figure 3 | Fabrication, structure and logic function of a PNNTA tile.

a, Schematic of key components of the two-block PNNTA tile, including assembled and patterned Ge/Si nanowires (cyan) with source and drain electrodes (blue), and charge-trapping trilayer gate dielectric (purple) and metal gate lines (grey). The fabricated structure consists of two blocks of NWFETs, block 1 (left) and block 2 (right). **b**, Distribution of V_{th} from 70 NWFET nodes in block 1 in the PNNTA tile. The blue and red bars represent the V_{th} values of devices in inactive and active states, respectively, with

circuits, we selected devices with average V_{th} values of 1.0 ± 0.4 and 3.7 ± 0.5 V for active and inactive states, respectively (Supplementary Fig. 4). Similarly, the chosen NWFET nodes in block 2 had V_{th} values of 1.4 ± 0.8 and 4.0 ± 0.3 V for active and inactive states, respectively. The distinction between V_{th} values for both states in both blocks of the PNNTA tile provide a relatively wide, ~ 2 -V, window for circuit operation.

The two-block PNNTA tile was initially programmed to function as a full adder, an important combinational circuit in the arithmetic logic unit in modern digital computers. Figure 3c illustrates the configuration of the one-bit full-adder logic circuit comprising two blocks with the output of block 1 (Fig. 3c, left-hand box) fed into block 2 (Fig. 3c, right-hand box) as input through external wiring. The programmed active node pattern (Fig. 3c, green dots) determines the circuit function, and in this case the outputs S and C_{out} represent the sum and carry-out of the summation of inputs $A + B + C$, respectively, with $S = A \oplus B \oplus C$ and $C_{out} = A \cdot B + A \cdot C + B \cdot C$. The symbols ' \oplus ', ' \cdot ' and ' $+$ ' represent logical XOR, AND and OR, respectively. Typical voltage transfer functions of the resulting circuit for power-supply voltage, V_{DD} , of 3.0 V (Fig. 3d) show that as the input levels of A , B and C are swept from logic state 0 (0 V) to logic state 1 (3.5 V), the outputs S and C_{out} switch from logic 0 (both 0 V) to logic 1 (2.0 and 2.7 V, respectively). From this data, the peak voltage gains of C_{out} and S (Fig. 3d, lines tangential to data) are found to be 10 and 4, respectively.

$V_{ds} = 0.5$ V. **c**, Circuit design implementing a one-bit full adder. I_A , I_B and I_C denote the complementary inputs of A , B and C , respectively. The left- and right-hand dashed boxes outline block 1 and block 2, respectively. **d**, Voltage transfer function for S (red) and C_{out} (blue) from input states $(0, 0, 0)$ to $(1, 1, 1)$. The dashed tangent lines show the maximal voltage gains of the outputs. **e**, Output voltage levels for S and C_{out} for six typical input states. **f**, Truth table of full-adder logic for the six input states in **e**. The measured output voltages are shown in brackets.

The larger-than-unity gain and the matching of input–output voltage levels are crucial for potentially cascading the logic tiles (Supplementary Fig. 2). Further tests showed that the output of S and C_{out} for six typical input combinations (Fig. 3e) all had similar output ranges: 0–0.6 V for logic state 0 and 2.0–2.7 V for logic state 1. The expected and experimental results for a full adder are summarized in a truth table (Fig. 3f), which shows good consistency for this fundamental logic unit. The V_{th} value of some active NWFET nodes shifted with the 3-V source bias and precluded switching behaviour for the $(A, B, C) = (0, 1, 0)$ and $(0, 1, 1)$ inputs for a consistent input voltage range (0–3.5 V). We note that optimization of logic operations can be achieved by tuning V_{DD} and the load resistance, together with adjustment of V_{th} through the choice of top-gate metal¹². Nonetheless, the large voltage gain and matching of input–output voltage levels described here show the potential to integrate the prototype device into large-scale integrated circuits such as a multi-bit adder in a cascade configuration.

Notably, the same PNNTA tile can be used to perform a range of distinct logic operations because we can reproducibly and independently reprogram the active and inactive nodes in both blocks (Methods and Supplementary Fig. 5). To illustrate this key point, we first reprogrammed the same tile shown in Fig. 3 to function as a full subtractor (Fig. 4a). The two outputs of the reprogrammed circuit, D and B_{out} ,

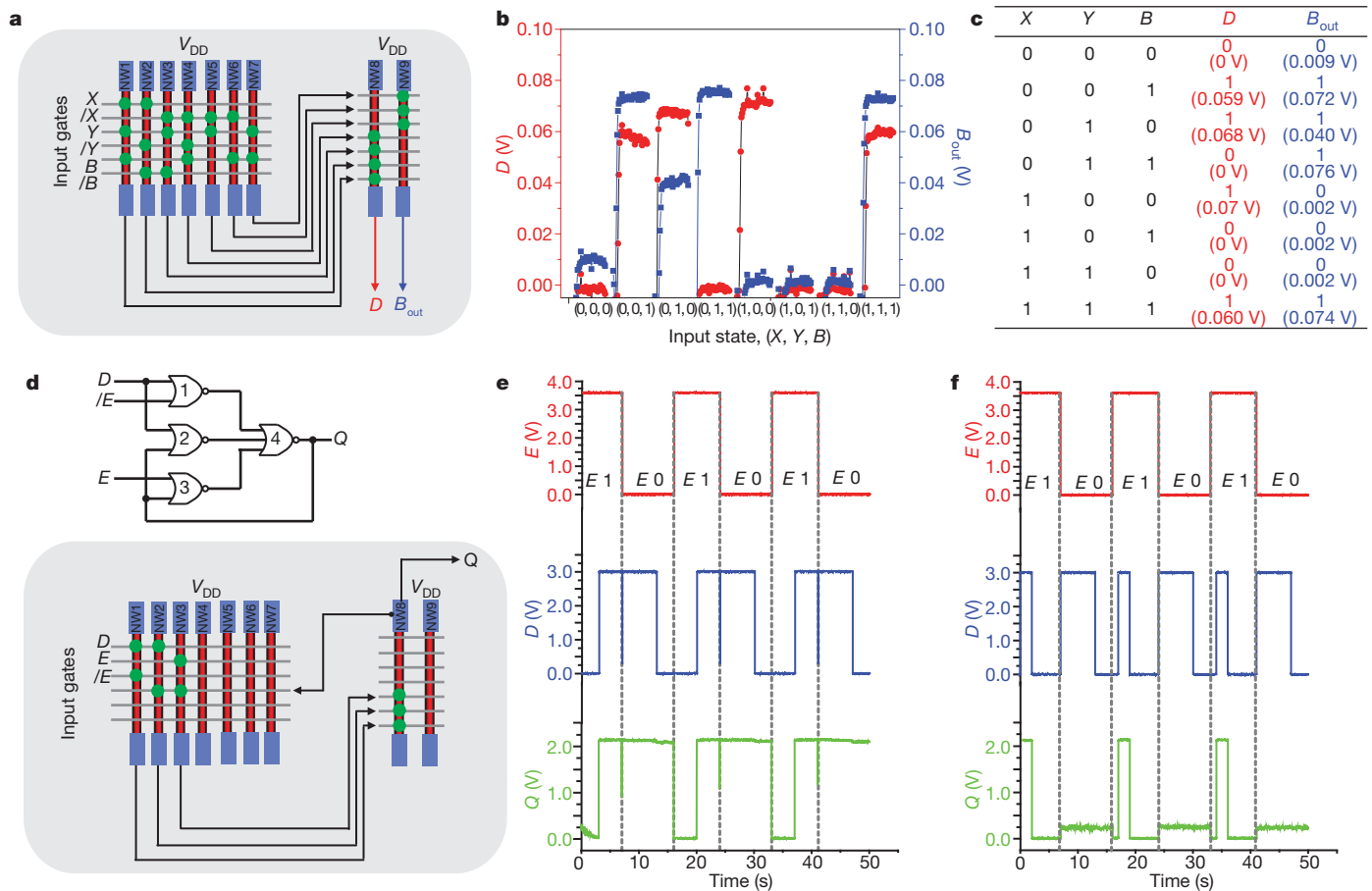


Figure 4 | Multifunctional PNNTA architecture. **a**, Schematic of a circuit implementing a full subtractor. **b**, Output of D (red) and B_{out} (blue) of the full subtractor implemented with the same PNNTA structure shown in Fig. 3 with eight input states. **c**, Truth table of the full subtractor with measured output

voltages shown in brackets. **d**, Schematics of logic (upper) and circuit design (lower) of a D latch implemented with the same PNNTA tile used in **a–c**. **e**, **f**, Output, Q , waveforms (green) at two sets of clock (E , red) and data (D , blue) inputs.

represent the difference and borrow, respectively, of the subtraction of inputs $X - Y - B$, with $D = X \oplus Y \oplus B$ and $B_{out} = B \cdot (\bar{X} \oplus Y) + \bar{X} \cdot Y$, where \bar{X} represents the logical negation of X (that is, the complementary input). Measurements of D and B_{out} for different $X - Y - B$ input combinations (Fig. 4b) show that the output voltage levels for logic state 0 (0–0.01 V) and logic state 1 (0.04–0.08 V) are well separated and represent robust states. Moreover, the truth table summarizing the expected and experimental results for the full subtractor (Fig. 4c) shows full and correct logic for this processing unit. In addition, we used the same tile to program and demonstrate multiplexer and demultiplexer circuits (Supplementary Fig. 6), showing the capability and flexibility of the PNNTA to fulfil the core functions of combinational circuit elements.

Significantly, we can also use our nanowire tile as a sequential circuit element, which represents another critical component beyond the scope of combinational elements. To do so, we mapped a D latch¹⁹, a sequential logic circuit capable of information storage, onto the unit tile (Fig. 4d). The D-latch circuit (Fig. 4d, upper panel) is composed of four NOR gates with a positive-feedback connection between the output, Q , and inputs to NOR gates 2 and 3 (Fig. 4d, upper panel). As a consequence, Q equals input data, D , when clock, E , is in logic state 1 but retains its previous value when E is switched into logic state 0. We implemented the NOR gates in the tile using NW1–NW3 in block 1 and NW8 in block 2 (Fig. 4d, lower panel), and formed the positive feedback by connecting the output to an input gate in block 1. An important constraint on realizing the D-latch logic is that there must be a successful feedback loop (output Q of block 2 back to block 1; Fig. 4d), which requires matching of input and output voltage levels.

Measurement of Q as a function of repetitive E and D pulses (Fig. 4e) shows that Q follows D when E is switched to logic 1 (3.6 V) at time points 16 and 33 s but retains its previous value when E is switched to logic 0 (0 V) at 7, 24 and 41 s, as expected for a D latch. The robustness of this sequential logic circuit was tested further by inputting a more complex data waveform (Fig. 4f), where measurements of Q demonstrated sharp logic operation by following D with high fidelity in the time intervals 16–24 and 33–41 s. Moreover, the voltage range of output, Q (0–2.2 V), closely matches that of input data, D , and clock, E .

Our nanowire logic tile has novel features in comparison with previous circuits based on bottom-up nanoscale elements^{1,2,4–9}. First, the architecture enables us to enhance by a factor of at least three the complexity of nanoelectronic circuits assembled from the bottom up (56 devices of two-block, coupled logic rather than ≤ 16 uncoupled devices in a single block in previous work^{4–6}), and correspondingly has led to circuits exceeding simple logic realized using nanowires^{4–6}, carbon nanotubes^{7,8} and memristors⁹. Second, our circuits show a maximal voltage gain of ten, which is comparable to previous reports on simple nanowire and carbon nanotube logic devices^{4–8} and represents a significant advantage over passive memristor devices⁹, where gain is ≤ 1 . Gain is crucial for signal restoration^{10,11} and makes the PNNTA architecture suitable for larger-scale processors. Third, the reversible programming of individual NWFET nodes in the tile provides great versatility, as shown by the combinational and sequential circuit elements reported above. Reconfigurable logic has been realized using memristor complementary metal–oxide–semiconductor (CMOS) hybrid circuits²⁰, where the microscale CMOS layer is responsible for logic operation and memristors are responsible for reconfigurable signal routing. Our

architecture, however, represents the first example of a system integrating nanoscale devices that combine both logic and programmability functions. These bottom-up nanowire circuits also have limitations in comparison with conventional CMOS circuits, although projections suggest that the density, speed and power consumption can be further improved for our array architecture (Supplementary Information).

In summary, we have demonstrated a programmable and scalable architecture based on a unit logic tile consisting of two interconnected, programmable, non-volatile nanowire transistor arrays. Each NWFET node in an array can be programmed to act as an active or an inactive transistor state, and by mapping different active-node patterns into the array, combinational and sequential logic functions including full adder, full subtractor, multiplexer, demultiplexer and D-latch can be realized with the same programmable tile. Cascading this unit logic tile into linear or tree-like interconnected arrays, which will be possible given the demonstrated gain and matched input–output voltage levels of NWFET devices, provides a promising bottom-up strategy for developing increasingly complex nanoprocessors with heterogeneous building blocks^{2,21}. In the near term, particularly promising for this architecture and the low-power devices it contains are simpler, tiny, application-specific nanoelectronic control processors³; such ‘nanocounters’ might make possible very small embedded electronic systems and new types of therapeutic device.

METHODS SUMMARY

We synthesized the Ge/Si core/shell nanowires using a nanocluster-catalysed methodology described previously¹². Growth of the charge-trapping gate dielectric shells by atomic-layer deposition was carried out in a vacuum system (Savannah-100, Cambridge NanoTech) at 200 °C, using trimethylaluminium, tetrakis (dimethylamino)zirconium and water as precursors. The three layers were deposited without interruptions in between. Standard EBL and thermal evaporation were used to form metal electrodes (Ni for source and drain and Cr–Au for top gate). We used the focused ion beam technique to prepare a cross-sectional sample of the NWFET device, and used lubricant-assisted contact printing to prepare axially aligned Ge/Si nanowire arrays. EBL and inductively coupled plasma reactive ion etching were used to pattern the nanowires. Electrical measurements were made with a computer-controlled, analogue input–output system (National Instruments). A custom-designed 96-pin probe card (Accuprobe) was used to access devices in the PNNTA array electrically.

Received 6 August; accepted 6 December 2010.

1. Lu, W. & Lieber, C. M. Nanoelectronics from the bottom up. *Nature Mater.* **6**, 841–850 (2007).
2. Lu, W., Xie, P. & Lieber, C. M. Nanowire transistor performance limits and applications. *IEEE Trans. Electron. Dev.* **55**, 2859–2876 (2008).
3. Das, S. *et al.* Designs for ultra-tiny, special-purpose nanoelectronic circuits. *IEEE Trans. Circuits Syst. Regul. Pap.* **54**, 2528–2540 (2007).
4. Cui, Y. & Lieber, C. M. Functional nanoscale electronic devices assembled using silicon nanowire building blocks. *Science* **291**, 851–853 (2001).

5. Huang, Y. *et al.* Logic gates and computation from assembled nanowire building blocks. *Science* **294**, 1313–1317 (2001).
6. Zhong, Z. H., Wang, D. L., Cui, Y., Bockrath, M. W. & Lieber, C. M. Nanowire crossbar arrays as address decoders for integrated nanosystems. *Science* **302**, 1377–1379 (2003).
7. Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. Logic circuits with carbon nanotube transistors. *Science* **294**, 1317–1320 (2001).
8. Javey, A. *et al.* High- κ dielectrics for advanced carbon-nanotube transistors and logic gates. *Nature Mater.* **1**, 241–246 (2002).
9. Borghetti, J. *et al.* ‘Memristive’ switches enable ‘stateful’ logic operations via material implication. *Nature* **464**, 873–876 (2010).
10. DeHon, A. Array-based architecture for FET-based, nanoscale electronics. *IEEE Trans. Nanotechnol.* **2**, 23–32 (2003).
11. Das, S., Rose, G. S., Ziegler, M. M., Picconatto, C. A. & Ellenbogen, J. C. Architectures and simulations for nanoprocessor systems integrated on the molecular scale. *Lect. Notes Phys.* **680**, 479–513 (2005).
12. Xiang, J. *et al.* Ge/Si nanowire heterostructures as high-performance field-effect transistors. *Nature* **441**, 489–493 (2006).
13. Liu, J., Wang, Q., Long, S. B., Zhang, M. H. & Liu, M. A metal/Al₂O₃/ZrO₂/SiO₂/Si (MAZOS) structure for high-performance non-volatile memory application. *Semicond. Sci. Technol.* **25**, 055013 (2010).
14. Sze, S. M. *Physics of Semiconductor Devices* 438–445 (Wiley, 1981).
15. Javey, A., Nam, S., Friedman, R. S., Yan, H. & Lieber, C. M. Layer-by-layer assembly of nanowires for three-dimensional, multifunctional electronics. *Nano Lett.* **7**, 773–777 (2007).
16. Snider, G., Kuekes, P. & Williams, R. S. CMOS-like logic in defective, nanoscale crossbars. *Nanotechnology* **15**, 881–891 (2004).
17. Whang, D., Jin, S., Wu, Y. & Lieber, C. M. Large-scale hierarchical organization of nanowire arrays for integrated nanosystems. *Nano Lett.* **3**, 1255–1259 (2003).
18. Sakamoto, W. *et al.* in *Proc. Electronic Devices Meeting 2009*, doi:10.1109/IEDM.2009.5424211 (IEEE, 2009).
19. Sedra, A. S. & Smith, K. C. *Microelectronics Circuits* 1014–1021 (Oxford Univ. Press, 2004).
20. Xia, Q. F. *et al.* Memristor-CMOS hybrid integrated circuits for reconfigurable logic. *Nano Lett.* **9**, 3640–3645 (2009).
21. Nam, S., Jiang, X. C., Xiong, Q. H., Ham, D. & Lieber, C. M. Vertically integrated, three-dimensional nanowire complementary metal-oxide-semiconductor circuits. *Proc. Natl Acad. Sci. USA* **16**, 21035–21038 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Bell and N. Antoniou for transmission electron microscopy sample preparation and imaging, Q. Qing for assistance with electrical measurements and J. L. Huang, X. Duan and X. Jiang for helpful discussions. C.M.L. acknowledges support from a National Security Science and Engineering Faculty Fellow award and a contract from the MITRE Corporation. S.D., J.F.K. and J.C.E. acknowledge support by the US government’s Nano-Enabled Technology Initiative and the MITRE Innovation Program.

Author Contributions C.M.L., J.C.E., S.D., H.Y., H.S.C. and S.N. designed the experiments. H.Y., H.S.C., S.N., Y.H. and J.F.K. performed the experiments. S.D. performed simulations. H.Y., H.S.C., S.N., S.D., J.F.K., J.C.E. and C.M.L. analysed the data and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.D. (sdas@mitre.org) or C.M.L. (cml@cmliris.harvard.edu).

Asymmetric additions to dienes catalysed by a dithiophosphoric acid

Nathan D. Shapiro^{1*}, Vivek Rauniyar^{1*}, Gregory L. Hamilton¹, Jeffrey Wu¹ & F. Dean Toste¹

Chiral Brønsted acids (proton donors) have been shown to facilitate a broad range of asymmetric chemical transformations under catalytic conditions without requiring additional toxic or expensive metals^{1–8}. Although the catalysts developed thus far are remarkably effective at activating polarized functional groups, it is not clear whether organic Brønsted acids can be used to catalyse highly enantioselective transformations of unactivated carbon–carbon multiple bonds. This deficiency persists despite the fact that racemic acid-catalysed ‘Markovnikov’ additions to alkenes are well known chemical transformations. Here we show that chiral dithiophosphoric acids can catalyse the intramolecular hydroamination and hydroarylation of dienes and allenes to generate heterocyclic products in exceptional yield and enantiomeric excess. We present a mechanistic hypothesis that involves the addition of the acid catalyst to the diene, followed by nucleophilic displacement of the resulting dithiophosphate intermediate; we also report mass spectroscopic and deuterium labelling studies in support of the proposed mechanism. The catalysts and concepts revealed in this study should prove applicable to other asymmetric functionalizations of unsaturated systems.

It has been known for over a century that strong Brønsted acids can catalyse the addition of alcohols and other protic nucleophiles to simple alkenes. The ability to predict the regioselectivity of these reactions is taught in every introductory organic chemistry course as Markovnikov’s rule. However, successful approaches to asymmetric variants have relied on metal catalysts rather than organic Brønsted acids, particularly in the area of amine addition reactions^{9–12}. Although metal-free Brønsted acids can catalyse additions to unactivated alkenes with yields comparable to those produced by the use of metals^{13–15}, the lone example of an attempted enantioselective variant of this reaction using a chiral acid resulted in poor selectivity (17% enantiomeric excess, e.e.)¹⁶. Although a number of structurally diverse strong Brønsted acid catalysts have been developed, the highly enantioselective reactions reported to date are restricted to the activation of an electrophilic carbon–heteroatom or heteroatom–heteroatom multiple bond, usually an imine or a carbonyl^{1–8}.

This unfortunate limitation can perhaps be explained by considering the different intermediates generated by protonation of an imine or carbonyl versus an alkene (Fig. 1a). Protonation of an imine or carbonyl generates a species that can hydrogen-bond with the conjugate base of the chiral Brønsted acid. This hydrogen bond serves as an anchor to keep the chiral information close to the reactive electrophile and also contributes to the molecular organization that favours one particular diastereomeric transition state. On the other hand, protonation of an alkene leads to a carbocation. Although the conjugate base of the chiral acid can still be held in proximity to the carbocation through electrostatic interactions, the lack of rigidity in this association presumably results in poor discrimination between the enantiotopic faces of the carbocation. In fact, a recent review on chiral Brønsted acid catalysis goes as far as to say that “The key to realizing enantioselective catalysis using a chiral Brønsted acid is the hydrogen-bonding interaction

between a protonated substrate and the chiral conjugate base”²³. Clearly, a conceptually different approach is needed to achieve the desired enantioselective additions to alkenes.

We considered that this problem could be overcome for nucleophilic additions to dienes by using a chiral Brønsted acid with a nucleophilic conjugate base that could form a covalent bond with the carbocation (Fig. 1b). In a second step, the nucleophile could displace the chiral leaving group in an S_N2’ reaction (displacement of an allylic leaving group by nucleophilic attack at the alkene). Because the chiral catalyst is directly bound to the substrate in the nucleophilic addition step, we hypothesized that this mechanistic scheme might facilitate a highly enantioselective transformation. Notably, two of the most important modes of organocatalysis, namely enamine and iminium catalysis, also take advantage of ‘covalent catalysis’ mechanisms¹⁷.

A challenge in implementing such a strategy is finding an acid that is strong enough to protonate an alkene, but which also possesses a nucleophilic conjugate base. We considered that dithiophosphoric acids might be ideal candidates to fulfil both criteria¹⁸. The increased polarizability of sulphur (2.90) versus oxygen (0.802) makes dithiophosphoric acids more acidic and nucleophilic than their oxygenated analogues^{19–21}. For the purpose of our desired reaction, it was encouraging to note that the addition of achiral dithiophosphoric acids to dienes is known to proceed efficiently with Markovnikov regioselectivity under radical-free conditions²². We suspected that the challenge in reaction development would therefore arise in achieving a highly selective reaction, especially given that the single previously reported reaction using a chiral dithiophosphoric acid catalyst proceeded with low diastereoselectivity and enantioselectivity (7:3 diastereomeric ratio (d.r.), 63% e.e.)²³.

Putting our idea into practice, we found that chiral dithiophosphoric acid **3a** catalysed the intramolecular hydroamination of diene **1** to form the desired pyrrolidine product **2** with excellent yield and moderate

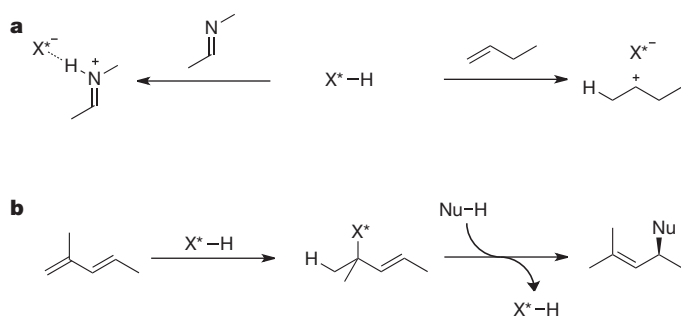
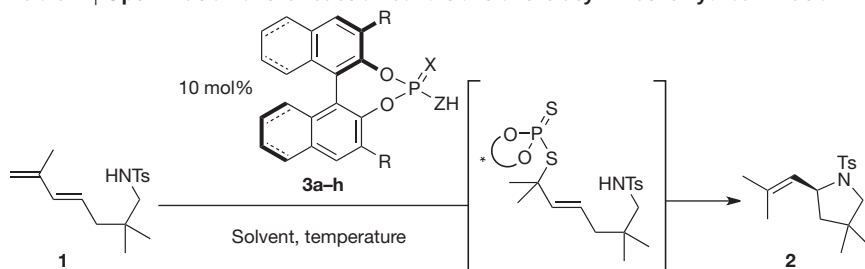


Figure 1 | A possible solution to the mechanistic challenge of asymmetric acid-catalysed additions to alkenes. **a**, Protonation of an imine with a chiral Brønsted acid (X*-H) leads to a hydrogen-bonded intermediate (left), while protonation of an alkene results in a carbocation (right) that cannot form a hydrogen bond. **b**, Proposed mechanism wherein a nucleophilic chiral acid adds to a diene then undergoes enantioselective S_N2’ displacement.

¹Department of Chemistry, University of California, Berkeley, California 94720, USA.

*These authors contributed equally to this work.

Table 1 | Optimization of the reaction conditions of the asymmetric hydroamination


Entry	Catalyst structure	Catalyst number	Solvent	Temperature (°C)	Yield (%)	e.e. (%)
1		3a X = Z = S, R = 1-naphthyl	CDCl ₃	30	91	41
2		3b X = Z = O, R = 1-naphthyl	CDCl ₃	30	0	NA
3		3c X = S, Z = NTf, R = 1-naphthyl	CDCl ₃	30	89	46
4		3d X = O, Z = NTf, R = 1-naphthyl	CDCl ₃	30	0	NA
5		3e R = 9-anthracenyl	CDCl ₃	30	98	62
6		3f R = 9-anthracenyl	CDCl ₃	15	91	78
7		3g R = 10-phenylanthracenyl	FC ₆ H ₅	15	92	94
8		3g R = 10-(3,5-bis- <i>t</i> -Bu-C ₆ H ₃)-9-anthracenyl	FC ₆ H ₅	15	96	96
9		3h R = 10-(2,4,6-(CH ₃) ₃ -C ₆ H ₂)-9-anthracenyl	FC ₆ H ₅	23	98	96

Ts, *p*-toluenesulfonyl; Tf, trifluoromethanesulfonyl. Reactions were all run for 48 h. Yields were determined by NMR analysis versus an internal standard; e.e.s were determined by chiral HPLC. NA, not available.

enantioselectivity (Table 1, entry 1). As expected, the oxygenated phosphoric acid analogue **3b** did not promote the reaction at all (entry 2). We also found that an *N*-trifluoromethanesulfonyl (*N*-triflyl) thiophosphoramidate catalyst of the type reported in ref. 6 catalysed the reaction with comparable e.e. (Table 1, entry 3), whereas the corresponding oxygen analogue **3d** did not give any desired product (entry 4). Attempts to optimize the catalyst structure by synthesizing more sterically encumbered *N*-triflyl thiophosphoramidates resulted in unacceptably low yields, so we continued our investigation with dithiophosphoric acids.

Changing the 3,3' substituents to bulkier anthracenyl groups led to a substantial boost in enantioselectivity, as did using a catalyst with a partially hydrogenated backbone (Table 1, entry 5). We also noted that performing the reaction in fluorobenzene as solvent in the presence of 4 Å molecular sieves at a slightly reduced temperature (15 °C) further improved the selectivity (Table 1, entry 6). Finally, based on the proposed S_N2'-type mechanism in which the incoming nucleophile is some distance away from the chiral dithiophosphate, we hypothesized that extending the catalyst structure could lead to even better results by more effectively 'projecting' the chiral information. Consistent with this proposal, addition of an aromatic substituent at the 10-position of the anthracene moiety allowed us to achieve excellent enantioselectivity (Table 1, entries 7–9). Notably, the mesityl catalyst **3h** provided exceptional enantioinduction even at room temperature. Because in some cases one catalyst offered slightly better selectivity than the other, we used both **3g** and **3h** for exploring the scope of the reaction.

A number of structural modifications could be made to the substrates while preserving the excellent yield and enantioselectivity of the catalytic hydroamination (Table 2). The sulfonyl group on the amine can be varied while maintaining the excellent yield and enantioselectivity of the reaction (Table 2, entries 1 and 2). The terminal alkene can also be freely substituted with cyclic or acyclic groups (Table 2, entries 3 and 4). Diene **4d** showed selectivity for the *E*-isomer of the product, although both geometric isomers were formed with high enantioselectivity and had the same absolute configuration at the newly formed stereogenic centre. Interestingly, complementary selectivity for the *Z*-alkene could be achieved by using the isomeric diene **4e** (Table 2, entry 5). In both cases, the major product was obtained in higher enantiomeric excess than the other alkene isomer. With regard to

functional group tolerance, it is remarkable to note that a primary *t*-butyldimethylsilyl (TBS) ether was stable in the presence of the strongly acidic catalyst in spite of the general acid lability of this protecting group (Table 2, entry 6). The tendency of the dithiophosphate to add covalently to the diene rather than remain free in solution may explain this surprising chemoselectivity. Additionally, the tether between the nucleophile and the diene can be varied to generate spirocyclic products (Table 2, entries 7 and 8).

In considering our mechanistic hypothesis, we realized that we should be able to access the same type of allylic dithiophosphate ester intermediate from addition of the Brønsted acid catalyst to allenes (1,2-dienes). We found that allene substrate **4i** was indeed converted to the pyrrolidine product **2** with essentially the same yield and enantioselectivity as was observed starting from the corresponding 1,3-diene (Table 2 entry 9, compare Table 1 entry 8). This observation also held true for other substrates. Although sulfonyl-pyrrolidines are themselves useful compounds from a medicinal chemistry standpoint^{24,25}, we also wanted to prepare products where the nitrogen substituent could be cleaved under mild conditions. Towards this end, we found that a 2-nitrosulfonyl (nosyl)-protected amine could be synthesized with only a modest decrease in enantioselectivity (Table 2, entry 10, 90% e.e.). Perhaps unsurprisingly, a more drastic change to a phosphinyl protecting group resulted in a slightly greater drop in selectivity (Table 2, entry 11). Hydroxylamines also proved to be useful substrates for the reaction, providing isoxazolidine products with very good enantioselectivities (Table 2, entries 13 and 14). Although in general we obtained the best results with substrates that possess geminal disubstitution in the alkyl tether, an observation probably attributable to the Thorpe–Ingold effect, the high enantioselectivity obtained using allene **4n** demonstrates that this is not strictly necessary for the success of our reaction.

A number of additional experiments were performed in order to further elucidate the mechanism of this transformation (Fig. 2). We began by analysing aliquots taken during the course of the catalytic reaction of **1** using time-of-flight mass spectrometry (TOF-MS). We observed a new peak that was fully consistent (mass-to-charge ratio *m/z* and isotopic distribution) with proposed intermediate **6** (Fig. 2a, Supplementary Figs 4 and 5). The proposed formation of this intermediate is also supported by the fact that the addition of dithiophosphoric acids across alkenes and dienes is a well-established process^{20,22,26,27}.

Table 2 | Performance of various 1,2- and 1,3-dienes in the enantioselective hydroamination reaction

Entry	Diene (4a–4n)	No.	Temperature (°C)	Product (5a–5n)	No.	Yield (%E:Z)	e.e. (%)
1		4a	23		5a	99	92
2		4b	23		5b	99	95
3		4c	30		5c	70	94
4		4d	30		5d	90 (4.7:1)	95 (E) 90 (Z)
5		4e	23		5d	75 (1:2)	91 (E) 99 (Z)
6		4f	23		5f	91 (1:3.6)	80 (E) 99 (Z)
7		4g	23		5g	99	96
8		4h	23		5h	91	97
9		4i	23		2	99	95
10		4j	23		5j	81	90
11		4k	23		5k	99	83
12		4l	40		5l	67	97
13		4m	23		5m	70	90
14		4n	60		5n	67	92

Reactions were run in fluorobenzene for 48 h using 10 mol% **3g** (entries 3 and 4) or 10 mol% **3h** (all others, 20 mol% for entry 14) in the presence of 4 Å molecular sieves. Yields refer to isolated material. TBS, *t*-butyldimethylsilyl; Ns, 2-nitrophenylsulphonyl.

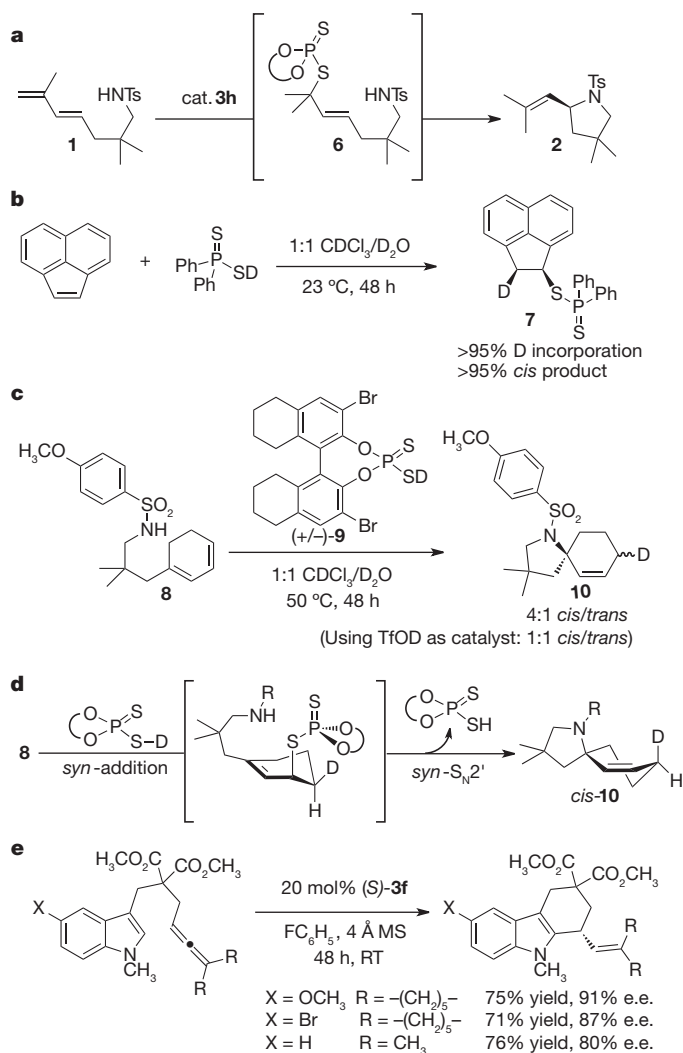


Figure 2 | Experiments to elucidate the reaction mechanism and application to indole nucleophiles. **a**, Proposed reaction mechanism involving a covalently bound catalyst–substrate intermediate that undergoes S_N2' displacement. Ts (tosyl), *p*-toluenesulfonyl. **b**, Addition of an achiral dithiophosphinic acid across an alkene proceeds with *syn* stereoselectivity. **c**, Reaction of a cyclic substrate using deuterated catalyst reveals 1,4-*syn*-stereoselectivity. Tf (triflyl), trifluoromethanesulfonyl. **d**, The overall mechanistic picture suggested by these experiments involves initial *syn*-addition of the S–H(D) bond across the alkene, followed by *syn*- S_N2' displacement. R, $\text{SO}_2(4\text{-CH}_3\text{O-C}_6\text{H}_4)$. **e**, Dithiophosphoric acid-catalysed hydroarylation of indole derivatives. MS, molecular sieve; RT, room temperature.

An investigation of the diastereoselectivity of the protonation and nucleophilic addition steps revealed some more insights regarding the mechanism. A deuterated achiral dithiophosphinic acid added across acenaphthylene, a cyclic alkene often used as a stereochemical probe, with a very high level of *syn*-stereoselectivity (Fig. 2b). No epimerization of the product was observed, even after a prolonged reaction time with heating (50 °C, 72 h). Thus, at least in this case, the dithiophosphinate ester intermediate does not ionize under conditions harsher than those used in the catalytic reaction. We next examined the reaction of a cyclic diene-tethered sulphonamide substrate using a deuterated catalyst (Fig. 2c). The obtained spirocyclic product was substantially enriched (4:1 d.r.) in the isomer where the sulphonamide nucleophile and the deuterium have a *cis* orientation. Taken together, these two experiments suggest that this observed *syn* diastereoselectivity is a result of initial *syn*-addition of the dithiophosphoric acid across the distal alkene, followed by a *syn*- S_N2' displacement (Fig. 2d). Excluding metal-mediated processes, S_N2' reactions are known to proceed preferentially through *syn* pathways^{28,29}.

At this point we cannot describe with certainty the degree of bonding that exists between the nucleophile, allylic system, and dithiophosphate in the S_N2' displacement step. This step may be concerted, or it may involve the formation of an allylic carbocation–dithiophosphate tight ion pair that is rapidly trapped by the tethered sulphonamide. In either mechanism, the remarkable feature is that the catalyst is able to mediate the attack of the nucleophile on the carbon electrophile with sufficient organization to greatly favour one diastereomeric transition state. In addition, it should be noted that cyclization of stereochemical probe **8** using catalytic deuterated triflic acid proceeds with no diastereoselectivity (Fig. 2c). This result strongly supports the notion that the dithiophosphoric acid catalysed reaction is mechanistically distinct from simple Brønsted acid catalysis.

To demonstrate the generality of this approach, we examined indoles as useful carbon nucleophiles that would be structurally and mechanistically distinct from the sulphonamides used in the rest of the study. Although a large number of efficient additions of indoles to imine and unsaturated carbonyl derivatives have been discovered, the proposed organocatalytic enantioselective hydroarylation of an unactivated carbon unsaturated system has not been demonstrated³⁰. When indole substrates were subjected to our reaction conditions, the hydroarylations proceeded readily to afford the tetrahydrocarbazole products in good to excellent enantiomeric excess (80–91% e.e.; Fig. 2e). An X-ray structure of a crystalline sample of the brominated derivative confirmed the structure and revealed the absolute configuration of the products (Supplementary Fig. 11 and Supplementary Table 1).

The high enantioselectivity of this carbon–carbon bond forming reaction is particularly striking because the *N*-alkylated indole substrates do not possess any apparent hydrogen-bond donors to assist in the catalyst–substrate organization. As previously mentioned, the presence of hydrogen-bonding functionality has been a signature of nearly all of the previously demonstrated chiral Brønsted acid catalysed reactions³. It is possible that in our system, the covalent attachment of the catalyst eliminates the need for the hydrogen-bonding that is typically required for reactions that proceed by an ion pair mechanism. We believe that the applicability of these catalysts and concepts to this different type of bond formation augurs well for the scope of future developments.

In spite of the remarkable developments in the field of asymmetric catalysis, there are still a great number of important transformations that are beyond the reach of current synthetic approaches. We have reported here a method using dithiophosphoric acids that enables metal-free catalytic asymmetric nucleophilic additions to all-carbon π -systems. In addition to serving as a useful means of obtaining valuable chiral hetero- and carbo-cyclic products, the reported hydroamination and hydroarylation reactions are fundamentally distinct from those reactions that have been previously achieved using chiral organocatalysts. Finally, we have presented experimental evidence that is most consistent with a unique covalent catalysis mechanism.

METHODS SUMMARY

General procedure: to a 1-dram screw-cap vial was added the diene or the allene substrate (0.1 mmol, 1.0 equiv) followed by the dithiophosphoric acid catalyst **3f**, **3g** or **3h** (0.01 mmol, 0.1 equiv) and activated 4 Å molecular sieves (20 mg). To the mixture was added fluorobenzene (0.5 ml) at room temperature. The vial was sealed and allowed to stand for 48 h at the indicated temperature. After the reaction was complete, the entire mixture was loaded onto silica gel and the product was eluted with EtOAc/hexanes. For complete experimental details, including procedures and full characterization (¹H and ¹³C NMR, high-resolution mass spectrometry) of all new compounds, see Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 July; accepted 3 December 2010.

1. Akiyama, T., Itoh, J., Yokota, K. & Fuchibe, K. Enantioselective Mannich-type reaction catalyzed by a chiral Brønsted acid. *Angew. Chem. Int. Edn* **43**, 1566–1568 (2004).

2. Uraguchi, D. & Terada, M. Chiral Brønsted acid-catalyzed direct Mannich reactions via electrophilic activation. *J. Am. Chem. Soc.* **126**, 5356–5357 (2004).
3. Terada, M. Binaphthol-derived phosphoric acid as a versatile catalyst for enantioselective carbon–carbon bond forming reactions. *Chem. Commun.* 4097–4112 (2008).
4. Akiyama, T. Stronger Brønsted acids. *Chem. Rev.* **107**, 5744–5758 (2007).
5. Nakashima, D. & Yamamoto, H. Design of chiral *N*-triflyl phosphoramidate as a strong chiral Brønsted acid and its application to asymmetric Diels–Alder reaction. *J. Am. Chem. Soc.* **128**, 9626–9627 (2006).
6. Cheon, C. H. & Yamamoto, H. A Brønsted acid catalyst for the enantioselective protonation reaction. *J. Am. Chem. Soc.* **130**, 9246–9247 (2008).
7. Hatano, M., Maki, T., Moriyama, K., Arinobe, M. & Ishihara, K. Pyridinium 1,1'-binaphthyl-2,2'-disulfonates as highly effective chiral Brønsted acid–base combined salt catalysts for enantioselective Mannich-type reaction. *J. Am. Chem. Soc.* **130**, 16858–16860 (2008).
8. García-García, P., Lay, F., García-García, P., Rabalakos, C. & List, B. A powerful chiral counteranion motif for asymmetric catalysis. *Angew. Chem. Int. Edn* **48**, 4363–4366 (2009).
9. Aillaud, I., Collin, J., Hannedouche, J. & Schulz, E. Asymmetric hydroamination of non-activated carbon–carbon multiple bonds. *Dalton Trans.* 5105–5118 (2007).
10. Hultzsich, K. C. Transition metal-catalyzed asymmetric hydroamination of alkenes (AHA). *Adv. Synth. Catal.* **347**, 367–391 (2005).
11. Hultzsich, K. C. Catalytic asymmetric hydroamination of non-activated olefins. *Org. Biomol. Chem.* **3**, 1819–1824 (2005).
12. Roesky, P. W. & Müller, T. E. Enantioselective catalytic hydroamination of alkenes. *Angew. Chem. Int. Edn* **42**, 2708–2710 (2003).
13. Schlummer, B. & Hartwig, J. F. Brønsted acid-catalyzed intramolecular hydroamination of protected alkenylamines. Synthesis of pyrrolidines and piperidines. *Org. Lett.* **4**, 1471–1474 (2002).
14. Rosenfeld, D. C., Shekhar, S., Takemiya, A., Utsunomiya, M. & Hartwig, J. F. Hydroamination and hydroalkoxylation catalyzed by triflic acid. Parallels to reactions initiated with metal triflates. *Org. Lett.* **8**, 4179–4182 (2006).
15. Li, Z. *et al.* Brønsted acid catalyzed addition of phenols, carboxylic acids, and tosylamides to simple olefins. *Org. Lett.* **8**, 4175–4178 (2006).
16. Ackermann, L. & Althammer, A. Phosphoric acid diesters as efficient catalysts for hydroaminations of nonactivated alkenes and an application to asymmetric hydroamination. *Synlett* 995–998 (2008).
17. MacMillan, D. W. C. The advent and development of organocatalysis. *Nature* **455**, 304–308 (2008).
18. Norman, G. R., LeSuer, W. M. & Mastin, T. W. Chemistry of the aliphatic esters of thiophosphoric acids. II. O,O,S-trialkyl thionophosphates by the addition of O,O-dialkyl thiolthionophosphoric acids to olefins. *J. Am. Chem. Soc.* **74**, 161–163 (1952).
19. Kabachnik, M. I., Mastrukova, T. A., Shipov, A. E. & Melentyeva, T. A. The application of the Hammett equation to the theory of tautomeric equilibrium: thione–thiol equilibrium, acidity, and structure of phosphorous thioacids. *Tetrahedron* **9**, 10–28 (1960).
20. Mastryukova, T. A., Spivak, L. L., Grigoreva, A. A., Urzhuntseva, E. K. & Kabachnik, M. I. Ionization constants of phosphorus dithioacids in absolute ethanol. *J. Gen. Chem. USSR (Engl. Transl.)* **41**, 1938–1941 (1971).
21. Leavitt, C. M. *et al.* Investigations of acidity and nucleophilicity of diphenyldithiophosphinate ligands using theory and gas-phase dissociation reactions. *Inorg. Chem.* **47**, 3056–3064 (2008).
22. Mueller, W. H. & Oswald, A. A. Organic sulfur compounds. XV. Cationic addition of O,O'-diethylthiophosphoric acid to olefins. *J. Org. Chem.* **31**, 1894–1898 (1966).
23. Pousse, G. *et al.* Synthesis of BINOL derived phosphorodithioic acids as new chiral Brønsted acids and an improved synthesis of 3,3'-disubstituted H₈-BINOL derivatives. *Tetrahedron* **65**, 10617–10622 (2009).
24. Cheng, X. C., Wang, Q., Fang, H., Tang, W. & Xu, W. F. Synthesis of new sulfonyl pyrrolidine derivatives as matrix metalloproteinase inhibitors. *Bioorg. Med. Chem.* **16**, 7932–7938 (2008).
25. Mutel, V. & Wichmann, J. Sulfonyl-pyrrolidine derivatives useful for the treatment of neurological disorders. European patent 1,299 383 (2005).
26. Obushak, N. D., Vovk, M. V., Vengrzhansovskii, V. A., Mel'nik, Y. I. & Ganushchak, N. I. Regioselective addition of O,O-diethyl dithiophosphate to 1-aryl-1,3-butadienes. *J. Gen. Chem. USSR (Engl. Transl.)* **57**, 1078–1080 (1987).
27. Mebah, J. M. N., Mieloszynski, J. L. & Paquer, D. Reactivity of O,O-diethylthiophosphoric acid toward different ethylenic systems. *Phosphorus Sulfur Silicon Relat. Elem.* **73**, 49–56 (1992).
28. Houk, K. N., Paddon-Row, M. N. & Rondan, N. G. The origin of syn stereoselectivity in the S_N2' reaction. *J. Mol. Struct.* **103**, 197–208 (1983).
29. Borrmann, T. & Stohrer, W. What governs the stereochemistry of the S_N2' reaction? *Liebigs Ann.* **1996**, 1593–1597 (1996).
30. Liu, C. & Widenhoefer, R. A. Gold(I)-catalyzed intramolecular enantioselective hydroarylation of allenes with indoles. *Org. Lett.* **9**, 1935–1938 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to D. Benitez and E. Tkatchouk (Goddard Group, California Institute of Technology) for performing DFT calculations, to U. Anderson and T. Iavarone of the QB3/Chemistry Mass Spectrometry Facility, and to A. DiPasquale of the X-ray Crystallography Facility. We acknowledge funding from the Director, Office of Science of the US Department of Energy (contract no. DE-AC02-05CH11231) and from NIH/GMS (R01 GM074774). V.R. is grateful for a Natural Sciences and Engineering Research Council of Canada (NSERC) postdoctoral fellowship. G.L.H. is grateful for a Bristol-Myers/Squibb graduate fellowship.

Author Contributions N.D.S. initiated the hydroamination study. V.R. optimized the catalysts and initiated the hydroarylation study. N.D.S., V.R., G.L.H. and J.W. performed the experiments. N.D.S., G.L.H. and F.D.T. developed the mechanistic concepts. G.L.H. and N.D.S. wrote the manuscript with input from all authors.

Author Information X-ray crystallographic data have been deposited in the Cambridge Crystallographic Data Centre database (<http://www.ccdc.cam.ac.uk/>) under code CCDC 800545. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.D.T. (fdtoste@berkeley.edu).

METHODS

General reaction procedure: to a 1-dram screw-cap vial was added the diene or the allene substrate (0.1 mmol, 1.0 equiv) followed by the dithiophosphoric acid catalyst **3f**, **3g** or **3h** (0.01 mmol, 0.1 equiv) and activated 4 Å molecular sieves (20 mg). To the mixture was added fluorobenzene (0.5 ml) at room temperature. The vial was sealed and allowed to stand for 48 h at the indicated temperature. After the reaction was complete, the entire mixture was loaded onto silica gel, and the product was eluted with ethyl acetate/hexanes. Unless otherwise noted, all commercial materials were used without further purification. Small-scale reactions were conducted in one-dram vials fitted with a threaded cap. All other reactions were conducted in flame-dried glassware under an N₂ atmosphere with magnetic stirring and dried solvent. Solvents were dried by passage through an activated

alumina column under argon. Thin-layer chromatography (TLC) analysis of reaction mixtures was performed using Merck silica gel 60 F254 TLC plates and visualized by a combination of ultraviolet and potassium permanganate staining. Flash column chromatography was carried out on Merck 60 silica gel (32–63 µm). Nuclear magnetic resonance (NMR) spectra were recorded with Bruker AV-600, AVB-400, AVQ-400 and AV-300 spectrometers. ¹H and ¹³C chemical shifts are reported in p.p.m. relative to tetramethylsilane, ³¹P chemical shifts are reported relative to 85% aqueous phosphoric acid, and ¹⁹F chemical shifts are reported relative to CFCl₃. Multiplicities are reported using the following abbreviations: s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; br, broad resonance. Enantiomeric excesses (e.e.s) were determined on a Shimadzu VP Series Chiral HPLC. Mass spectral data were obtained by the QB3/Chemistry Mass Spectrometry Facility at UC Berkeley.

Holocene Southern Ocean surface temperature variability west of the Antarctic Peninsula

A. E. Shevenell^{1†}, A. E. Ingalls¹, E. W. Domack² & C. Kelly^{1‡}

The disintegration of ice shelves, reduced sea-ice and glacier extent, and shifting ecological zones observed around Antarctica^{1,2} highlight the impact of recent atmospheric³ and oceanic warming⁴ on the cryosphere. Observations^{1,2} and models^{5,6} suggest that oceanic and atmospheric temperature variations at Antarctica's margins affect global cryosphere stability, ocean circulation, sea levels and carbon cycling. In particular, recent climate changes on the Antarctic Peninsula have been dramatic, yet the Holocene climate variability of this region is largely unknown, limiting our ability to evaluate ongoing changes within the context of historical variability and underlying forcing mechanisms. Here we show that surface ocean temperatures at the continental margin of the western Antarctic Peninsula cooled by 3–4 °C over the past 12,000 years, tracking the Holocene decline of local (65° S) spring insolation. Our results, based on TEX₈₆ sea surface temperature (SST) proxy evidence from a marine sediment core, indicate the importance of regional summer duration as a driver of Antarctic seasonal sea-ice fluctuations⁷. On millennial timescales, abrupt SST fluctuations of 2–4 °C coincide with globally recognized climate variability⁸. Similarities between our SSTs, Southern Hemisphere westerly wind reconstructions⁹ and El Niño/Southern Oscillation variability¹⁰ indicate that present climate teleconnections between the tropical Pacific Ocean and the western Antarctic Peninsula¹¹ strengthened late in the Holocene epoch. We conclude that during the Holocene, Southern Ocean temperatures at the western Antarctic Peninsula margin were tied to changes in the position of the westerlies, which have a critical role in global carbon cycling^{9,12}.

The Antarctic Peninsula is the northernmost extension of Antarctica (Fig. 1a). Its maritime climate results from the influence of the Southern Hemisphere westerlies and the Antarctic Circumpolar Current (ACC), which transmit climate variations to and from lower latitudes^{2,13}. Ongoing western Antarctic Peninsula (WAP) warming (3.4 °C per century) and associated environmental change^{1–3} highlight regional climate sensitivity and may result from a southward migration or intensification of the westerlies and/or the ACC^{4,14}. Modern regional SSTs (0–150 m; 1993–2004¹³) have an average annual range of –1.5 to 2 °C and reflect changes in solar insolation, wind forcing and sea-ice dynamics (Fig. 1c)^{2,13}. Ocean–atmosphere heat exchange occurs over the WAP shelf¹³. Warm (1–2 °C) Circumpolar Deep Water (CDW), partially derived from North Atlantic Deep Water, upwells onto the shelf (>200 m) owing to local bathymetry and the ACC's proximity, and is entrained in surface waters through mixing and local topographic upwelling¹³ (Fig. 1). A recent increase in CDW upwelling along the WAP is associated with winter SSTs greater than 0 °C, reduced sea-ice extent and duration, increased cyclogenesis, and El Niño/Southern Oscillation (ENSO) variability^{2,13,14}.

Although Antarctic ice cores provide details of past near-surface temperature^{15,16}, past Antarctic-margin SSTs are unknown because U^k₃₇, oxygen isotope and/or Mg/Ca palaeothermometry are not possible owing to the absence of haptophyte-derived alkenones and planktonic

foraminiferal CaCO₃ in regional marine sediments (Supplementary Information). Consequently, an emerging organic palaeothermometer, TEX₈₆ (the tetraether index of tetraethers with 86 carbon atoms; see Methods)¹⁷, based on the relative distribution of membrane lipids from pelagic marine archaea (glycerol dialkyl glycerol tetraethers (GDGTs)), holds regional appeal. Pelagic marine archaea are abundant in WAP surface waters, their regional ecology is studied and archaeal GDGTs are present in marine sediments¹⁸ (Fig. 2a and Supplementary Fig. 1). The premise of TEX₈₆ is that archaea adjust their membrane lipid composition in response to temperature and that sedimentary GDGTs are derived from the remains of surface-dwelling marine archaea¹⁷. TEX₈₆ values are converted to SST (0–150 m) using core-top calibrations that continue to evolve as new data emerge^{17,19} (Methods and Supplementary Information).

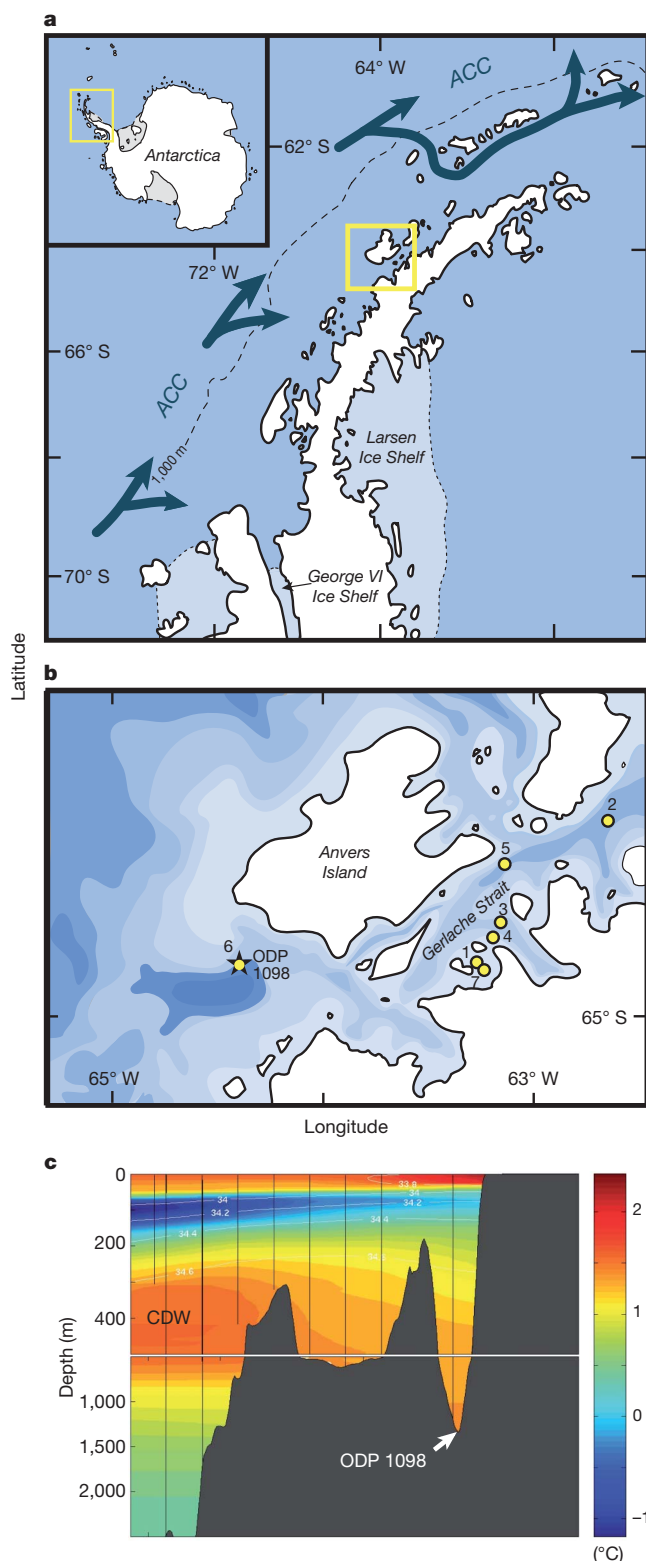
To test the regional utility of TEX₈₆ palaeothermometry, document Holocene SSTs and investigate processes forcing regional SSTs on orbital and millennial timescales, we generated a decadal–centennial-resolution, TEX₈₆-based SST record from a 43-m-thick Holocene sediment sequence at ODP Site 1098 (Hole 1098B; 64° 51.162' S, 64° 12.4795' W; 1,010-m water depth; Fig. 1; see Methods)²⁰. Existing palaeontological, sedimentological and geochemical studies at Site 1098 reveal millennial-scale palaeoclimate sensitivity^{21–23} (Fig. 2), but none provide robust SST estimates.

We converted TEX₈₆ values from ODP Hole 1098B to SST using a modified calibration that integrates new regional core-top data, with a global calibration data set¹⁹ (Fig. 1b; see Methods and Supplementary Information). Resulting SSTs range from –2.9 to 10.5 °C, with a mean of 1.67 °C ($n = 125$; pooled s.d. of 20 replicates, ± 0.8 °C (9%); Fig. 2). SSTs below –2 °C ($n = 3$) result from low TEX₈₆ values and are unrealistic considering the freezing point (–1.7 °C) of regional surface waters¹³. To determine whether our choice of calibration influences SST trends, we compared TEX₈₆-derived temperatures from several existing calibrations^{17,19} (Fig. 2a; see Supplementary Information). Regardless of the calibration or method of estimating error, all curves have similar trends, suggesting that although the absolute values of reported SSTs may change as calibrations and knowledge of archaeal ecology evolve, the observed trends are robust (Methods and Supplementary Information).

SSTs from ODP Site 1098 show orbital-scale Holocene (12–2 kyr ago) cooling ($(3–4) \pm 2.2$ °C; Fig. 3d). This cooling trend agrees with independent WAP palaeoclimate records (see ref. 23 for a review; also see Supplementary Information), further supporting the regional utility of TEX₈₆ palaeothermometry. Relatively high (average, 3.7 °C) early-Holocene (11.8–9 kyr ago) SSTs may reflect intermittent meltwater-induced stratification related to deglaciation. Today, SSTs greater than 6 °C have been measured in regional fjords during meltwater-induced stratification events. A similar process could explain the magnitude and variability of early Holocene SSTs at Site 1098, when annual insolation²⁴ and meltwater input²¹ were higher than present. Increased CDW influence may also explain this variability²³.

¹School of Oceanography, University of Washington, Seattle, Washington 98195, USA. ²Department of Geosciences, Hamilton College, Clinton, New York 13323, USA. [†]Present address: Departments of Geography and Earth Sciences, University College London, Gower Street, London WC1E 6BT, UK.

[‡]Deceased.



Superimposed on the Holocene cooling trend are a series of multi-centennial- to millennial-scale SST variations (2–4 °C; Fig. 4d). Warm events occurred ~11, 9, 6.5, 5.5, 3.5 and 1 kyr before present (BP), and cold events ~7.5, 6, 4.5, 2.3 and 0.2 kyr BP (Fig. 4d). Independent geologic evidence indicates that the George VI Ice Shelf (Fig. 1a) collapsed 9.6 kyr BP²³, following 2000 yr of relatively high early-Holocene SSTs at Site 1098^{15,16} (Fig. 3d). The ice shelf reformed and advanced at 7.9 kyr BP²³ as SSTs decreased (Fig. 4d). The lowest Holocene SSTs occurred between 2.7 and 1.7 kyr BP (average, -1.1 °C), coincident with regional glacial advances, increased sea-ice extent and reduced

Figure 1 | Western Antarctic Peninsula study location and oceanography. **a**, Antarctica (inset) and the Antarctic Peninsula with the study region outlined with a yellow box. Arrows indicate Antarctic Circumpolar Current (ACC) flow and the dashed line approximates the edge of the continental shelf (modified from ref. 23). **b**, Regional bathymetry with the Ocean Drilling Program (ODP) Site 1098 (star) and surface sediment sample locations (numbered yellow circles; Supplementary Table 1) marked. **c**, Austral summer (January 2001) cross-section of ocean temperatures at Site 1098; Palmer LTER Line 600 (<http://oceaninformatics.ucsd.edu/datazoo/data/pallter/studies>)¹³. In the upper ~100 m, SSTs reflect seasonal changes in insolation, atmospheric circulation, and meltwater influx^{2,13}. Warm (1–2 °C) Circumpolar Deep Water¹³ fills the basin below 200 m.

lacustrine productivity²³. A pronounced late-Holocene (1,600 to 500 yr BP) warming (average SST, 2.9 °C) terminated orbital-scale cooling (Fig. 2a). Terrestrial evidence for reduced glacial ice extent on nearby Anvers Island²⁵ (970 to 700 yr BP) corroborates observed WAP warmth. SST decrease between 500 and 200 yr BP agrees with evidence for increased regional sea ice, glacial and ice-shelf advances, and a decline in penguin populations²³. TEX₈₆ palaeothermometry indicates that regional SSTs increased by 3.2 °C in the past ~100 yr (Fig. 2c; see Methods), which is comparable to the observed warming trend, of 3.4 °C per century^{1–3}.

On orbital timescales, SSTs from ODP Site 1098 agree with Antarctic/sub-Antarctic palaeotemperature reconstructions (Fig. 3). Antarctic ice-core records indicate early-Holocene (11.5–9 kyr BP) warming^{15,16} (1–2.5 °C; Fig. 3c). In the Ross Sea sector, early-Holocene warmth is followed by a cooling trend of 2 °C (ref. 16), which is echoed in South Pacific (40–50° S) SSTs^{26,27} (Fig. 3c, e, f). Although Holocene cooling is widespread, its magnitude (3–4 °C) is greatest at ODP Site 1098 (Fig. 3d) and similar to modelled spring cooling trends in the seasonal sea-ice zone⁶ (2.5–3.5 °C). Thus, orbital-scale Holocene surface cooling at Site 1098 may be exaggerated by proxy seasonality, ice–albedo feedbacks and/or local oceanographic conditions.

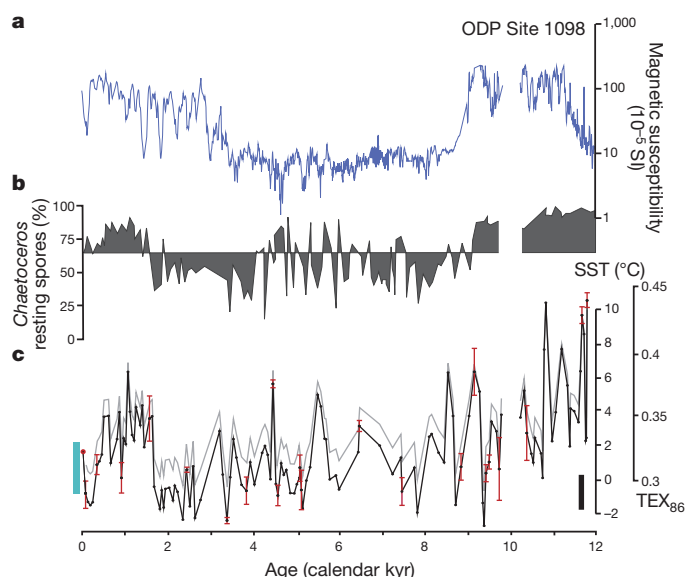


Figure 2 | Magnetic susceptibility²¹, abundance of *Chaetoceros* resting spores²² and TEX₈₆-derived SSTs versus calendar age, at ODP Site 1098. **a**, Magnetic susceptibility²¹ (SI, SI volume unit); **b**, abundance of *Chaetoceros* resting spores, shaded around average value²²; **c**, SSTs plotted using regional (black) and linear¹⁷ calibrations (grey). TEX₈₆ values are indicated. The red dot at far left marks the average regional core-top SST using regional calibration (Supplementary Table 1). Twenty replicates were analysed: pooled s.d., ±0.8 °C (analytical error; red error bars, 1σ). Black bar reflects total error: ±2.2 °C. Aqua bar indicates average annual SST range (1993–2004, -1.5–2 °C) at Site 1098; located in Palmer LTER Line 600¹³. Data gap at ~10 kyr represents a turbidite²⁰.

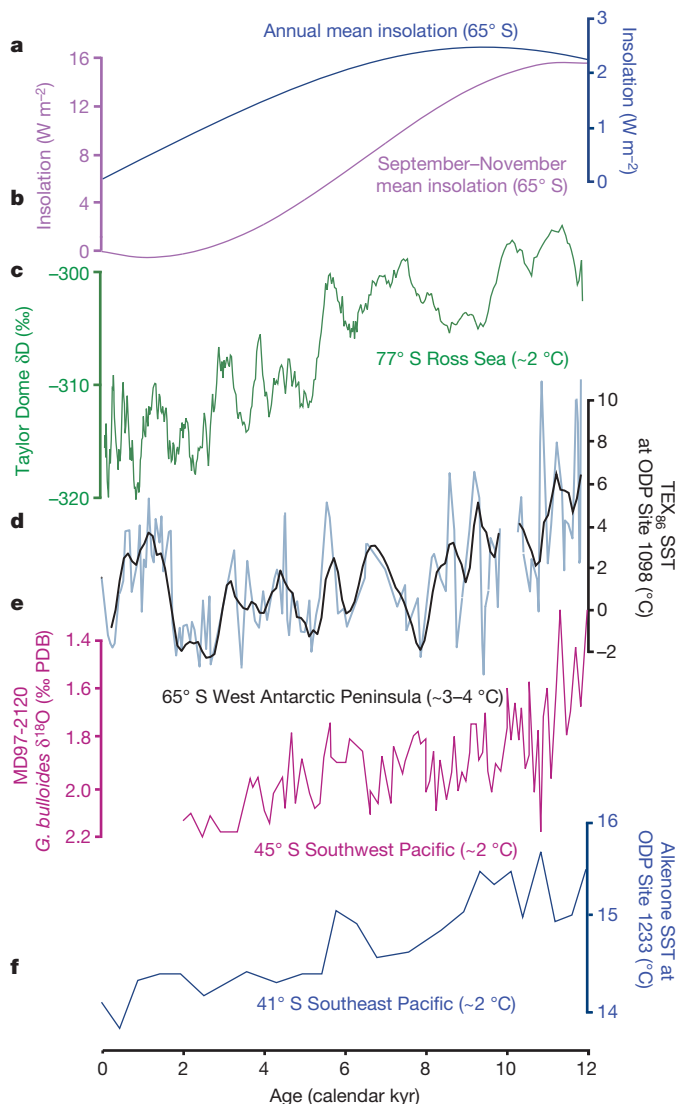


Figure 3 | Orbital-scale Holocene SST trends at ODP Site 1098 compared with insolation, Antarctic ice-core and sub-Antarctic Pacific Ocean SST records. **a, b**, Annual mean insolation at 65° S (navy; **a**) and austral spring (September–November) mean insolation at 65° S (lavender; plotted with reference season set to day 261 (ref. 6); **b**), plotted as deviations from present-day mean²⁴. **c**, Taylor Dome hydrogen isotope deuterium ($\delta D = (HD^{16}O/H_2^{16}O/R_{VSMOW} - 1) \times 1,000\text{‰}$, where R_{VSMOW} is the ratio of $HD^{16}O$ to $H_2^{16}O$ for Vienna Standard Mean Ocean Water; ref. 16; green) and **d**, TEX₈₆-derived SSTs at ODP Site 1098 (blue; black, five-point smoothing) reveal a long-term decrease in West Antarctic atmospheric and surface ocean temperatures of ~2–4 °C. **e, f**, Planktonic foraminifer (*Globigerina bulloides*) $\delta^{18}O$ ($((^{18}O/^{16}O)_{\text{sample}}/(^{18}O/^{16}O)_{\text{standard}} - 1) \times 1,000\text{‰}$, expressed relative to Vienna Pee Dee Belemnite standard; ref. 27; red; **e**) and alkenone²⁶ (blue, **f**) SST estimates indicate cooling of ~2 °C in the sub-Antarctic Pacific.

Existing regional data suggest that pelagic marine archaea are most abundant during the austral spring sea-ice retreat¹⁸. Thus, TEX₈₆-derived SSTs should be sensitive to local (65° S) spring insolation, which influences both the timing of regional sea-ice retreat² and the summer duration⁷. Orbital-scale Holocene cooling is consistent with declining austral spring insolation intensity at 65° S (ref. 24; Fig. 3). Our interpretation is strengthened by model results, which suggest local insolation forcing of Southern Ocean SSTs, sea-ice extent and the location of the westerlies⁶.

Proxy seasonality and local insolation forcing may explain the discrepancy in maximum Holocene warmth between our SST record, derived from early-spring pelagic marine archaea populations, and diatom abundances that reflect average summer conditions and

indicate mid-Holocene warmth²². Early-Holocene insolation at 65° S was higher (+16 W m⁻²) than at present in the late winter/spring (Fig. 3a, b), whereas late spring/early summer insolation reached an optimum 6 kyr BP^{6,24} (+9 W m⁻²). Although semi-quantitative, the greatest TEX₈₆ GDGT abundances occur during early-Holocene (12–10 kyr BP) and late-Holocene (1.6–0.5 kyr BP) warmth (Supplementary Information), coincident with high *Chaetoceros* resting spore abundances²² (Fig. 2b). Along the WAP, *Chaetoceros* diatoms bloom in early spring. The abundance of resting spores during warm intervals suggests elevated early-spring productivity²². As most pelagic marine archaea live as chemoautotrophs that use ammonia from decaying phytoplankton for energy²⁸, the correspondence between *Chaetoceros* resting spores and GDGTs suggests that archaeal production and subsequent GDGT export to sediments may be tightly coupled to the regional spring phytoplankton bloom.

It has been proposed⁷ that Southern Hemisphere summer duration, which covaries with Northern Hemisphere insolation intensity, regulates Antarctic temperatures on orbital timescales. It is argued that precession-driven changes in austral spring insolation control Antarctic summer duration, with implications for global climate related to the extent and duration of Antarctic sea ice⁷. Our results suggest that orbital-scale Holocene cooling resulted from a decline in high-latitude summer duration, amplified by regional ice–atmosphere–ocean feedbacks (for example sea-ice expansion, northern migration of the westerlies and reduced CDW upwelling). Our findings suggest that invoking climatic linkages with the Northern Hemisphere high latitudes is not necessary on orbital timescales^{16,29} and instead support hypotheses favouring Antarctic climate independence³⁰.

Millennial-scale SST variability (Fig. 4), which cannot be explained by orbital forcing, leaves open the possibility that Antarctic-margin SSTs were forced remotely through changes in atmospheric–ocean circulation. Within the limits of the respective age models (± 100 yr), millennial-scale SST variations at ODP Site 1098 are synchronous and in phase with temperature events recognized in Pacific-sector Antarctic ice cores^{15,16} and Chilean margin/southwest Pacific SSTs^{9,26} (Fig. 4). Beyond the Antarctic/sub-Antarctic, a series of six global, millennial-scale climate events are recognized⁸. SSTs from Site 1098 show fluctuations during five of these events (Fig. 4).

The near-synchronous millennial-scale response to global events favours atmospheric forcing of WAP SST by means of the westerlies, which may vary as a result of local processes or remote climate teleconnections¹⁴. Today, periods of regional cyclogenesis increase CDW upwelling on the WAP shelf, resulting in early spring sea-ice retreat^{2,13}. We suggest that millennial-scale Holocene warm and cold events at ODP Site 1098 resulted, respectively, from increased and decreased westerly wind strength/influence. In the former case this enhanced CDW upwelling and led to early sea-ice retreat and in the latter case it reduced CDW upwelling and led to late sea-ice retreat. Similarities between SSTs at Site 1098 and westerly wind reconstructions from South America⁹ support this hypothesis. Further support comes from Siple Dome melt layers¹⁵, which indicate that the maximum pre-anthropogenic Holocene warmth/maritime influence occurred between 1,600 and 500 kyr BP.

Today, WAP atmospheric and oceanic temperatures and sea-ice extent show the strongest response to ENSO of any region outside the tropical Pacific^{2,11}. Along the WAP, relatively high SSTs and reduced sea-ice extent coincide with cold La Niña events² owing to the formation of a high-pressure system in the Bellingshausen Sea, which brings warm air from lower latitudes towards Antarctica via the westerlies¹¹. Positive feedbacks in the seasonal sea-ice zone enhance regional warming^{2,11}. We propose that climate teleconnections between the tropical Pacific and the WAP strengthened in the late Holocene (~2 kyr BP; Fig. 4). The abrupt late-Holocene warming at ODP Site 1098 coincides with maximum Holocene ENSO variability¹⁰, pronounced La Niña-like conditions in the Tropical Pacific³¹ and maximum pre-anthropogenic Holocene atmospheric CO₂ concentrations³²

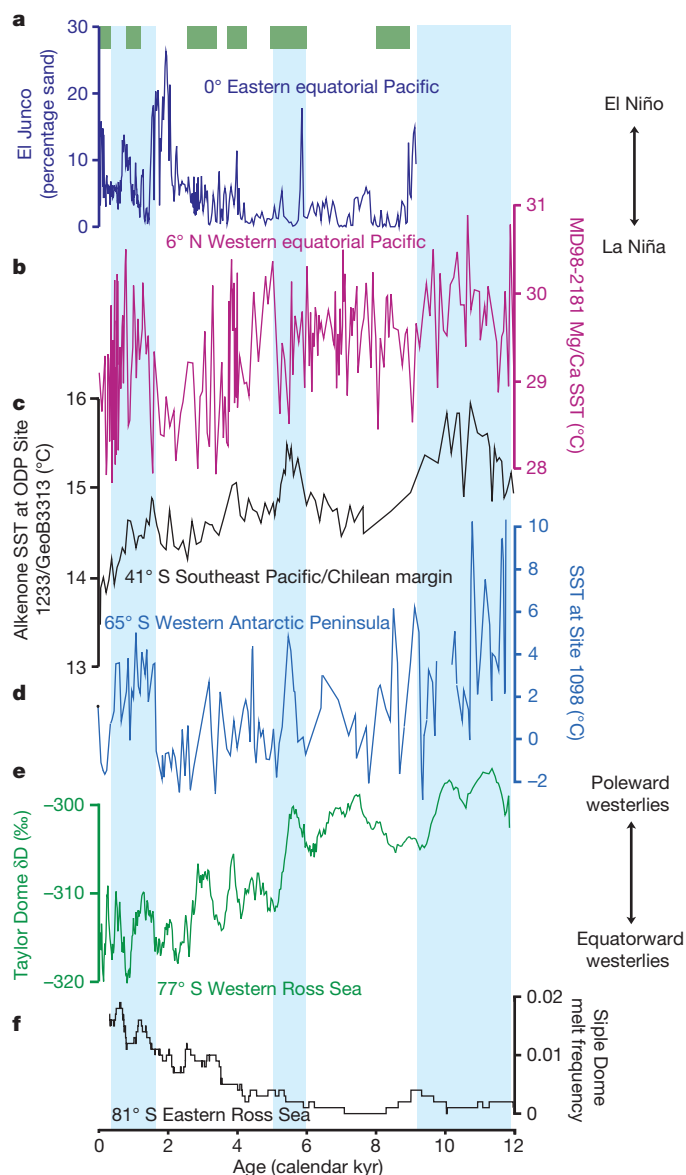


Figure 4 | Millennial-scale Holocene SST variability at ODP Site 1098 compared with Ross-Sea-sector Antarctic ice-core records, southeastern and western equatorial Pacific SSTs and Holocene ENSO frequency. **a**, Sand abundance at El Junco lake, Galapagos, reflects Holocene ENSO frequency¹⁰. **b**, Western equatorial Pacific SSTs³¹. **c**, Composite southeast Pacific/Chilean margin alkenone SSTs²⁶. **d**, TEX₈₆-derived SSTs at Site 1098. **e**, Taylor Dome δD (ref. 16). **f**, Siple Dome melt-layer frequency (1-kyr running mean); increased frequency corresponds to warmer temperatures and maritime influences¹⁵. Blue shading highlights millennial-scale warmings (see text). Green boxes denote the global millennial-scale intervals of ref. 8.

(Fig. 4). If ENSO increases in strength and frequency, as predicted for future climatic warming, this teleconnection may strengthen, with implications for Antarctic ice-sheet stability, ocean–atmosphere exchange of heat and CO₂, and sea-level rise.

Models and observations attribute the warming of the WAP by 3.4 °C in the twentieth century to anthropogenically forced climate change that has altered the strength and/or the position of the westerlies^{1–4,14}. On longer timescales, variations in Southern Hemisphere westerlies are linked to glacial–interglacial carbon cycling^{9,12}, suggesting that ongoing anthropogenic forcing of westerly winds may alter the Southern Ocean carbon sink. Our TEX₈₆-derived SST record from ODP Site 1098 provides a centennial–millennial-scale perspective on recent WAP warming, suggesting that late-Holocene warmth is in opposition to long-term Holocene climate trends and may relate to changes in the strength and/

or the position of the westerlies. Our results suggest that TEX₈₆-derived Antarctic-margin SST records, such as that from Site 1098, should advance our understanding of climate forcings and feedbacks at the ocean–cryosphere interface and improve ice-sheet model parameterizations⁵. Detailed WAP studies of the last 2 kyr are required to confirm the origin of warmth observed at Site 1098/Anvers Island²⁵ and document twentieth-century warming with greater temporal resolution than is possible at Site 1098.

METHODS SUMMARY

ODP Site 1098's chronology, previously established from 51 radiocarbon analyses of acid-insoluble organic matter and foraminiferal calcite, provides a calibrated timescale from 0 to 13 kyr BP²¹. We ultrasonically extracted freeze-dried sediment samples (~0.5 g) three times each with methanol, dichloromethane and methanol–dichloromethane (1:1), combined them and dried them under N₂. GDGTs used to generate the TEX₈₆ index were separated with an Agilent 1100-series high-performance liquid chromatograph and detected by atmospheric-pressure chemical ionization mass spectrometry using an Agilent XCT ion trap mass spectrometer at the University of Washington. We calculated the TEX₈₆ index using manually integrated extracted ion chromatograms of the appropriate peak areas and the formula of ref. 17. To determine the regional applicability of TEX₈₆ palaeothermometry, we analysed seven surface sediment samples (0–2 cm, corresponding to ~10 yr) for GDGTs; SSTs were measured during collection (spring–summer).

We used published calibrations to convert core-top TEX₈₆ values to SSTs and then compared calculated and measured SSTs (Supplementary Table 1). Using the linear equation of ref. 17 and a linear equation incorporating the complete data set of ref. 19, TEX₈₆-derived SSTs are within 2.5 °C of measured temperatures and close to the standard error, of 2 °C, of published calibrations (Supplementary Table 1). TEX₈₆-derived SSTs have the same trends as measured temperatures, where sediments underlying colder waters gave lower TEX₈₆-derived SSTs, although the temperature range is larger in the calculated values than in the measured data. Down-core TEX₈₆ values at Site 1098 were converted to temperature using a linear calibration that combines our seven regional core tops with the complete –2–30 °C data set of ref. 19 (Supplementary Fig. 2): $SST (^{\circ}C) = (0.0125 \times TEX_{86}) + 0.3038$ ($r^2 = 0.8237$) (Supplementary Information). We favour this approach because the calibration successfully estimates modern SSTs to within a total error of ± 2.2 °C, incorporates a large global TEX₈₆ calibration data set, and yields realistic SSTs at the cold end of the calibration (for example above the regional freezing point of sea water).

Received 10 August; accepted 10 December 2010.

1. Cook, A. J., Fox, A. J., Vaughan, D. G. & Ferrigno, J. G. Retreating glacier fronts on the Antarctic Peninsula over the past half-century. *Science* **308**, 541–544 (2005).
2. Stammerjohn, S. E., Martinson, D. G., Smith, R. C. & Iannuzzi, R. A. Sea ice in the western Antarctic Peninsula region: spatio-temporal variability from ecological and climate change perspectives. *Deep-Sea Res.* **55**, 2041–2058 (2008).
3. Steig, E. J. *et al.* Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. *Nature* **457**, 459–462 (2009).
4. Gille, S. T. Warming of the Southern Ocean since the 1950s. *Science* **295**, 1275–1277 (2002).
5. Pollard, D. & DeConto, R. M. Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458**, 329–332 (2009).
6. Renssen, H. *et al.* Holocene climate evolution in the high-latitude Southern Hemisphere simulated by a coupled atmosphere–sea ice–ocean–vegetation model. *Holocene* **15**, 951–964 (2005).
7. Huybers, P. & Denton, G. Antarctic temperature at orbital timescales controlled by local summer duration. *Nature Geosci.* **1**, 787–792 (2008).
8. Mayewski, P. A. *et al.* Holocene climate variability. *Quat. Res.* **62**, 243–255 (2004).
9. Moreno, P. I., Francois, J. P., Moy, C. M. & Villa-Martinez, R. Covariability of the Southern Westerlies and atmospheric CO₂ during the Holocene. *Geology* **38**, 727–730 (2010).
10. Conroy, J. L., Overpeck, J. T., Cole, J. E., Shanahan, T. M. & Steinitz-Kannan, M. Holocene changes in eastern tropical Pacific climate inferred from a Galapagos lake sediment record. *Quat. Sci. Rev.* **27**, 1166–1180 (2008).
11. Yuan, X. ENSO-related impacts on Antarctic sea ice: a synthesis of phenomenon and mechanisms. *Antarct. Sci.* **16**, 415–425 (2004).
12. Anderson, R. F. *et al.* Wind-driven upwelling in the Southern Ocean and the deglacial rise in atmospheric CO₂. *Science* **323**, 1443–1448 (2009).
13. Martinson, D. G. *et al.* Western Antarctic Peninsula physical oceanography and spatio-temporal variability. *Deep-Sea Res.* **55**, 1964–1987 (2008).
14. Lubin, D., Wittenmyer, R. A., Bromwich, D. H. & Marshall, G. J. Antarctic Peninsula mesoscale cyclone variability and climatic impacts influenced by the SAM. *Geophys. Res. Lett.* **35**, 1–4 (2008).
15. Das, S. B. & Alley, R. B. Rise in frequency of surface melting at Siple Dome through the Holocene: evidence for increasing marine influence on the climate of West Antarctica. *J. Geophys. Res.* **113**, 1–11 (2008).

16. Steig, E. J. *et al.* Synchronous climate changes in Antarctica and the North Atlantic. *Science* **282**, 92–95 (1998).
17. Schouten, S., Hopmans, E. C., Schefuß, E. & Sinninghe Damsté, J. S. Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth Planet. Sci. Lett.* **204**, 265–274 (2002).
18. Murray, A. E. *et al.* Seasonal and spatial variability of bacterial and archaeal assemblages in the coastal waters near Anvers Island, Antarctica. *Appl. Environ. Microbiol.* **64**, 2585–2595 (1998).
19. Kim, J.-H. *et al.* Global sediment core-top calibration of the TEX₈₆ paleothermometer in the ocean. *Geochim. Cosmochim. Acta* **72**, 1154–1173 (2008).
20. Shipboard Scientific Party in Proc. ODP, Initial Reps Vol. 178 (eds Barker, P. F. *et al.*) Ch. 7 (Ocean Drilling Program, 1999).
21. Domack, E. *et al.* Chronology of the Palmer Deep site, Antarctic Peninsula: a Holocene paleoenvironmental reference for the circum-Antarctic. *Holocene* **11**, 1–9 (2001).
22. Sjunneskog, C. & Taylor, F. Postglacial marine diatom record of the Palmer Deep, Antarctic Peninsula (ODP Leg 178, Site 1098) 1. Total diatom abundance. *Paleoceanography* **17**, 8003 (2002).
23. Bentley, M. J. *et al.* Mechanisms of Holocene palaeoenvironmental change in the Antarctic Peninsula region. *Holocene* **19**, 51–69 (2009).
24. Laskar, J. A long-term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428**, 261–285 (2004).
25. Hall, B. L., Koffman, T. & Denton, G. H. Reduced ice extent on the western Antarctic Peninsula at 700–970 cal yr B.P. *Geology* **38**, 635–638 (2010).
26. Kaiser, J., Schefuß, E., Lamy, F., Mohtadi, M. & Hebbeln, D. Glacial to Holocene changes in sea surface temperature and coastal vegetation in north central Chile: high versus low latitude forcing. *Quat. Sci. Rev.* **27**, 2064–2075 (2008).
27. Pahnke, K., Zahn, R., Elderfield, H. & Schulz, M. 340,000-year centennial-scale marine record of Southern Hemisphere climatic oscillation. *Science* **301**, 948–952 (2003).
28. Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543–546 (2005).
29. Nielsen, S. H. H., Koç, N. & Crosta, X. Holocene climate in the Atlantic sector of the Southern Ocean: controlled by insolation or oceanic circulation? *Geology* **32**, 317–320 (2004).
30. Wolff, E. W., Fischer, H. & Rothlisberger, R. Glacial terminations as southern warmings without northern control. *Nature Geosci.* **2**, 206–209 (2009).
31. Stott, L. *et al.* Decline of surface temperature and salinity in the western tropical Pacific Ocean in the Holocene epoch. *Nature* **431**, 56–59 (2004).
32. Monnin, E. *et al.* Evidence for substantial accumulation rate variability in Antarctica during the Holocene, through synchronization of CO₂ in the Taylor Dome, Dome C and DML ice cores. *Earth Planet. Sci. Lett.* **224**, 45–54 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Murray and K. Kryc for the ODP Hole 1098B samples, F. Lamy for the ODP Site 1233/GEOB3313 alkenone data set, S. Emerson, K. Kreutz, R. Dunbar, M. Maslin, R. Anderson, A. Newton, A. Pearson and A. Tudhope for discussions, and J. Sachs, O. Kawka and L. Truxal for assistance with alkenone measurements. This research used samples collected by the Ocean Drilling Program and the United States Antarctic Program. This research was supported by NSF grants OPP-0620099 (A.E.S. and A.E.I.) and OPP-0732467 (E.W.D.).

Author Contributions A.E.S. and A.E.I. designed and contributed equally to the study. A.E.S. analysed the TEX₈₆ and isotope data (in the laboratory of J.P. Kennett) and wrote the paper. A.E.I. analysed the alkenone samples. E.W.D. provided surface sediment samples and regional climate expertise. C.K. prepared the samples. A.E.S., A.E.I. and E.W.D. discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.E.S. (a.shevenell@ucl.ac.uk).

Acoelomorph flatworms are deuterostomes related to *Xenoturbella*

Hervé Philippe¹, Henner Brinkmann¹, Richard R. Copley², Leonid L. Moroz³, Hiroaki Nakano^{4†}, Albert J. Poustka⁵, Andreas Wallberg⁶, Kevin J. Peterson⁷ & Maximilian J. Telford⁸

Xenoturbellida and *Acoelomorpha* are marine worms with contentious ancestry. Both were originally associated with the flatworms (Platyhelminthes), but molecular data have revised their phylogenetic positions, generally linking *Xenoturbellida* to the deuterostomes^{1,2} and positioning the *Acoelomorpha* as the most basally branching bilaterian group(s)^{3–6}. Recent phylogenomic data suggested that *Xenoturbellida* and *Acoelomorpha* are sister taxa and together constitute an early branch of Bilateria⁷. Here we assemble three independent data sets—mitochondrial genes, a phylogenomic data set of 38,330 amino-acid positions and new microRNA (miRNA) complements—and show that the position of *Acoelomorpha* is strongly affected by a long-branch attraction (LBA) artefact. When we minimize LBA we find consistent support for a position of both *acoelomorphs* and *Xenoturbella* within the deuterostomes. The most likely phylogeny links *Xenoturbella* and *Acoelomorpha* in a clade we call *Xenacoelomorpha*. The *Xenacoelomorpha* is the sister group of the *Ambulacraria* (hemichordates and echinoderms). We show that analyses of miRNA complements⁸ have been affected by character loss in the acoels and that both groups possess one miRNA and the gene *Rsb66* otherwise specific to deuterostomes. In addition, *Xenoturbella* shares one miRNA with the ambulacrarians, and two with the acoels. This phylogeny makes sense of the shared characteristics of *Xenoturbellida* and *Acoelomorpha*, such as ciliary ultrastructure and diffuse nervous system, and implies the loss of various deuterostome characters in the *Xenacoelomorpha* including coelomic cavities, through gut and gill slits.

In contrast to previous results^{1,2,4,9,10} (Fig. 1a), two recent phylogenomic studies have suggested a sister group relationship between *Acoelomorpha* and *Xenoturbella*. These studies disagree over where this clade might be placed, either at the base of Bilateria⁷ (Fig. 1b) or with the deuterostomes¹¹ (Fig. 1c). The *acoelomorph* genes studied, however, show extremely high rates of sequence evolution. This bias could result in susceptibility to the LBA artefact: a systematic error that may be compounded by the short internal branches around the origin of the Bilateria¹². Overcoming this potential artefact requires the analysis of large molecular data sets comprising many species and using a complex model of sequence evolution designed to reduce the impact of systematic errors¹².

We assembled a largely complete set of mitochondrial protein sequences from four acoels using expressed sequence tag (EST) databases. Better-fitting models of molecular evolution are expected to be less sensitive to systematic errors, and cross validation¹³ shows that the CAT model with a general time reversible (GTR) exchange rate matrix and gamma correction (CAT + GTR + Γ) fits best, followed by GTR + Γ then CAT + Γ . In the phylogeny inferred with the best model, acoels are the sister-group of *Xenoturbella* (posterior probability (PP) of 0.99) within deuterostomes (PP = 0.99) (Fig. 2). However, the

relationship between chordates, ambulacrarians and the *Xenoturbella*/acoel group is unresolved (PP = 0.47). A notable feature of Fig. 2 is the extremely fast evolutionary rate of acoels, which are nevertheless grouped with the slow-evolving *Xenoturbella*.

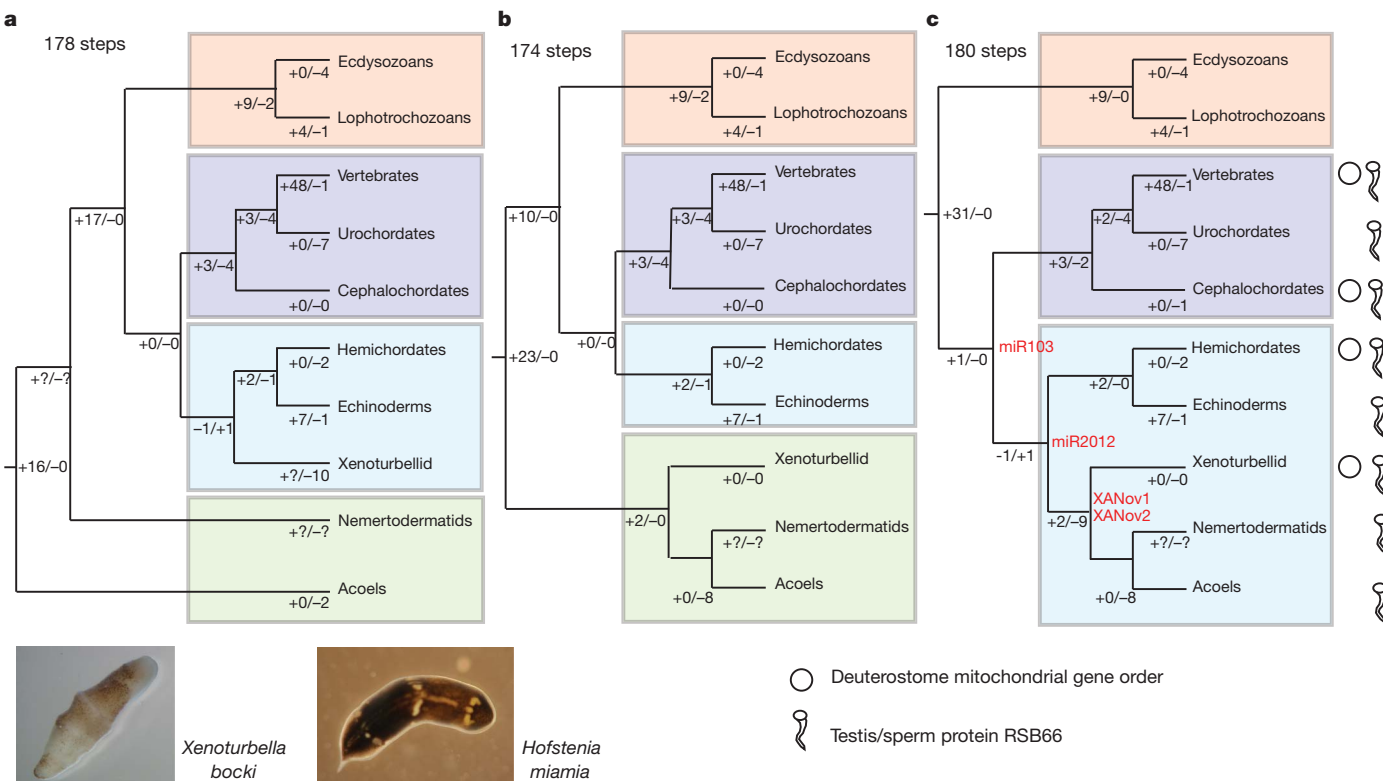
We exaggerated the effect of LBA by using the less-fit models (Supplementary Fig. 1). Using GTR + Γ we recover acoels + *Xenoturbella* (PP = 1.0) as basal deuterostomes (PP = 0.99). Only using the least fit model (CAT + Γ) do we find that the acoels are located as basal bilaterians (PP = 0.65). The fast evolutionary rate of acoels is therefore likely to be responsible for their early emergence revealed in previous studies¹⁴. Interestingly, the acoel *Symsagittifera roscoffensis* does not possess a protostome-type mitochondrial *NAD5* gene¹⁵, finally ruling out a close relationship between acoels and rhabditophoran platyhelminthes³ (Supplementary Fig. 2). Despite the limited size and heterogeneous evolutionary rates of mitochondrial genomes, these analyses provide evidence for grouping acoels and *Xenoturbella* together with deuterostomes.

We also constructed a large alignment from EST and genome data (66 species, 197 genes, 38,330 positions, 30% missing data) including all major animal phyla represented by slowly evolving species (Supplementary Tables 1 and 2). For this new data set, CAT + Γ has a better fit than GTR + Γ (CAT + GTR + Γ was not investigated as the computation is too time-consuming)¹³. The phylogeny inferred under CAT + Γ (Fig. 3) recovers all expected clades (Bilateria, Ecdysozoa, Lophotrochozoa, etc.) with high support (generally a bootstrap support (BS) of 100%). Acoels are seen to be very fast evolving and are the sister group of the nemertodermatids (BS = 55%). As in the mitochondrial analyses, the *acoelomorph* clade is sister to the slow-evolving *Xenoturbella* (BS = 80%). *Xenoturbella* plus *Acoelomorpha* are sister to *Ambulacraria* (BS = 78%) within deuterostomes.

Although the analysis of ESTs is congruent with the mitochondrial genome result, our topology differs from the recent phylogenomic analysis of Hejnol *et al.*⁷ (Fig. 1b). To test the possibility that the fast evolutionary rate of *Acoelomorpha* might have an effect on phylogenetic inference due to LBA, we pruned our data set to 37 species and compared alternative models (including CAT + GTR + Γ) and different taxon sampling schemes aimed at lessening or exaggerating a potential LBA artefact. The backbone topology inferred with the CAT + Γ model is unchanged when the number of taxa is reduced (Supplementary Fig. 3a and Fig. 3).

Cross-validation demonstrates that the site-heterogeneous CAT + GTR + Γ model has a significantly better fit than the CAT + Γ model ($\Delta\ln L = 490 \pm 48$), which itself is significantly better than the GTR + Γ model ($\Delta\ln L = 3,195 \pm 127$). Regardless of the species sampling, the best available model (CAT + GTR + Γ) locates *Xenoturbella*, *Acoela* and *Nemertodermatida* within deuterostomes. The fast-evolving *Nemertodermatida* are consistently found as a

¹Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada. ²Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. ³The Whitney Laboratory for Marine Biosciences and Department of Neuroscience, University of Florida, St Augustine and Gainesville, Florida 32080, USA. ⁴Department of Marine Ecology - Kristineberg, University of Gothenburg, 450 34 Fiskebäckskil, Sweden. ⁵Evolution and Development Group, Max-Planck Institut für Molekulare Genetik, Ihnestrasse 73, 14195 Berlin, Germany. ⁶Department of Systematic Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, Uppsala, Sweden. ⁷Department of Biology, Dartmouth College, Hanover, New Hampshire 03755, USA. ⁸Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK. [†]Present address: Shimoda Marine Research Center, University of Tsukuba, Shimoda, Shizuoka, 415-0025, Japan.



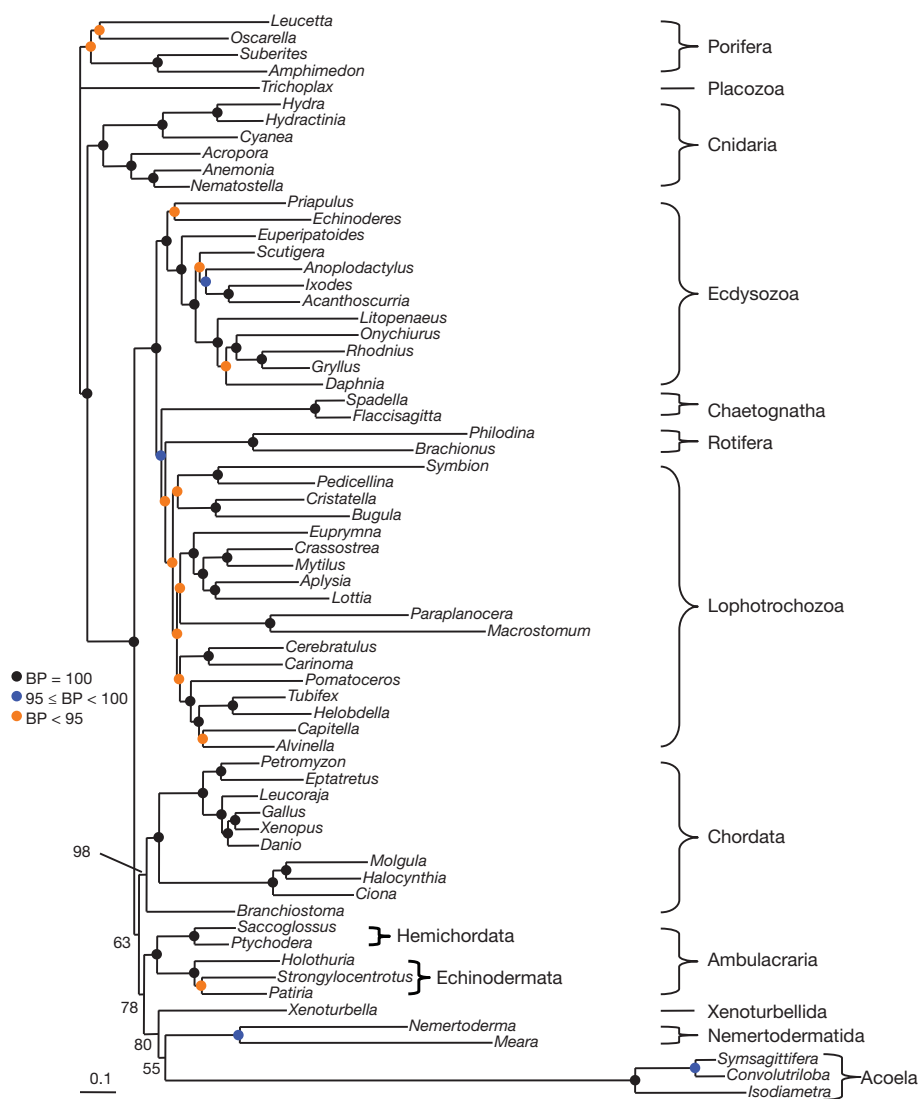


Figure 3 | Phylogeny of 66 animal species based on EST sequences. Analysis of 197 genes, 38,330 unambiguously aligned positions, 30% missing data. Tree was reconstructed using the CAT + Γ model under a Bayesian analysis. Major accepted metazoan clades (for example, Lophotrochozoa, Ecdysozoa, Protostomia) are supported. Acoela and Nemertodermatida are sister groups (Acoelomorpha). *Xenoturbella* and Acoelomorpha are sister groups (phylum Xenacoelomorpha). Xenacoelomorpha is the sister taxon of Ambulacraria (Xenambulacraria) within the deuterostomes. Level of bootstrap support is indicated. Similar support is obtained when jackknifing 50% of the genes (Supplementary Fig. 9). Scale bar, substitutions per position.

Computational constraints prevented the analysis of the original data set (94 species and 270,580 positions, with 84% missing data) with the site-heterogeneous CAT + Γ model; we therefore assembled an alignment of 145 genes (with the same gene coverage of the pivotal *Xenoturbella* and Acoelomorpha as used by Hejnal *et al.*) for the same 94 species (24,633 positions, and 30% missing data). The resulting phylogeny (Supplementary Fig. 6) is very similar to our results, with Nemertodermatida the sister-group to *Xenoturbella* (PP = 1.0), this clade being sister to Ambulacraria (PP = 0.98); the very fast-evolving acoels are included in deuterostomes (PP = 0.89), but with an unstable position (PP = 0.61 at the base of deuterostomes). Given that CAT + Γ has a better fit on this data set than the site-homogeneous model previously used, this suggests that the topology of the analysis of Hejnal *et al.*⁷ was affected by an LBA artefact.

Although our two data sets are consistent with a deuterostome affinity for Acoelomorpha and *Xenoturbella* (see also Supplementary Figs 7–9), the paucity of bilaterian miRNAs in one acoel, *Symsagittifera roscoffensis*, has supported the idea that the acoels are a basal clade relative to other Bilateria⁸. To examine this conclusion we have constructed and sequenced libraries of small RNAs from a second acoel, *Hofstenia miamia* and from *Xenoturbella bocki*. From the *Hofstenia* library we found ten miRNAs that were not detected in the *S. roscoffensis* library (Supplementary Table 3). From *Xenoturbella* we detected reads from all ten of these miRNAs, as well as eight additional miRNAs found in Bilateria. *Xenoturbella* has all but ten of the miRNAs typically found in bilaterian genomes¹⁶.

The most parsimonious tree derived from an analysis of miRNA data places the two species of acoels and *Xenoturbella* as three independent branches basal to the Bilateria (*Symsagittifera* (*Hofstenia* (*Xenoturbella*, Bilateria))) (Supplementary Figs 10 and 11). This result rather implausibly implies non-monophyly of Acoela. Alternative interpretations of these data assuming monophyly of acoels and based either on the results of refs 1, 2, 6, 9, 14 (Supplementary Fig. 12), or the tree of Hejnal *et al.* (Supplementary Fig. 13) or on our tree (Supplementary Fig. 14) imply large-scale losses of miRNAs, in particular from *Symsagittifera*. Numerous miRNAs must have been lost from at least some Acoels, suggesting that their absence cannot be considered a credible contra-indication of deuterostome affinity, fitting a picture of miRNA evolution occurring through continuous addition and mosaic loss¹⁶. Locating Acoelomorpha and *Xenoturbella* inside Deuterostomia, yet outside Chordata or Ambulacraria, means almost all possible losses are of bilaterian level characters—there is only a single known deuterostome specific miRNA—and this is exactly what we observe.

Limited additional support for our tree comes from this unique deuterostomian miRNA (miR-103/107/2013), which we find in both acoels and in *Xenoturbella* (Fig. 1 and Supplementary Fig. 14). *Xenoturbella*, at least, possesses a second miRNA, miR-1212 (Fig. 1 and Supplementary Fig. 15), previously found only in Ambulacraria. We suggest that miR-103/107/2013 is a plausible synapomorphy of *Xenoturbella*, Acoelomorpha, Ambulacraria and Chordata and that miR-1212 is a likely synapomorphy of Xenambulacraria. Furthermore, we find that *Xenoturbella* and acoels share two miRNAs found in no other animals: a

novel miRNA family (XANov-1) and a paralogue of miR-92 (XANov-2) (Supplementary Fig. 15). Finally, we have detected an additional gene, coding for the sperm protein RSB66, uniquely in the genomes of Ambulacraria, Chordata, Acoelomorpha and *Xenoturbella* (Fig. 1 and Supplementary Fig. 16). The *Rsb66* genes of acoelomorphs and *Xenoturbella* share a small and rather variable insertion relative to Chordata and Ambulacraria; this, in addition to their two novel miRNAs and their eight shared miRNA losses, gives further support to the idea that they are sister taxa (Fig. 1 and Supplementary Fig. 14).

Difficult phylogenetic questions such as that addressed here must ultimately be solved by the congruent patterns emerging from what, inevitably, are not highly supported results. Our three independent data sources indicate a sister group relationship between the acoelomorphs and *Xenoturbella*^{17–19} within the deuterostomes; we propose the name Xenacoelomorpha for this clade, noting that a deuterostome affinity for both *Xenoturbella* and Acoelomorpha has been previously suggested based on morphological considerations^{20–22}. The Xenacoelomorpha are excluded from the deuterostome phyla of Hemichordata, Echinodermata and Chordata and hence constitute an independent fourth phylum of deuterostomes. Our results suggest that characters shared by the Xenacoelomorpha are likely to be synapomorphies inherited from a common ancestor^{19,23}.

Our findings also indicate that the Acoelomorpha are not early branches on the stem leading to the Bilateria. This phylogenetic relationship, first reported over a decade ago⁶, had led to the acoelomorphs being interpreted as modern representatives of a lineage intermediate between the diploblasts and the Bilateria^{24,25}, a position that made sense of the paucity of *HOX* genes and miRNAs found in their genomes. The supposed presence of a small, simple, directly developing ancestor of the Bilateria with a weakly centralized nervous system and blind gut also led to the assumption—not supported by our findings—that these characteristics were present in Precambrian bilaterians²⁶.

Finally, the deuterostome affinity of the Xenacoelomorpha implies that they have lost characters present in the common ancestor of deuterostomes. This ancestor must have possessed pan-bilaterian apomorphies (for example, through-gut and protonephridia) as well as the homologous attributes of Ambulacraria and Chordata (deuterostomy, enterocoely, gill slits and endostylar tissue²⁷). Although it is clear that certain of these characters have been lost in the living Xenacoelomorpha, we predict that more deuterostome characters will be discovered in the morphology, embryology or genomes of members of the Xenacoelomorpha.

METHODS SUMMARY

Phylogenetic data. Data sets were gathered from National Center for Biotechnology Information databases. Data sets were assembled and aligned as in ref. 28 and analysed with the most complex site-heterogeneous model CAT + Γ and CAT + GTR + Γ^{29} as described in Supplementary Information.

miRNA data. Small RNA libraries were constructed and sequences analysed as described elsewhere³⁰. MicroRNA presequences were also recovered from *Xenoturbella* genomic DNA traces by BLAST searches. More details can be found in Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 August; accepted 16 November 2010.

1. Boursat, S. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85–88 (2006).
2. Boursat, S., Nielsen, C., Lockyer, A., Littlewood, D. T. J. & Telford, M. J. *Xenoturbella* is a deuterostome that eats molluscs. *Nature* **424**, 925–928 (2003).
3. Egger, B. *et al.* To be or not to be a flatworm: the acoel controversy. *PLoS ONE* **4**, e5502 (2009).
4. Telford, M. J., Lockyer, A. E., Cartwright-Finch, C. & Littlewood, D. T. J. Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. *Proc. R. Soc. Lond. B* **270**, 1077–1083 (2003).
5. Sempere, L. F., Cole, C. N., McPeck, M. A. & Peterson, K. J. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J. Exp. Zool.* **306**, 575–588 (2006).
6. Ruiz Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A. & Baguña, J. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science* **283**, 1919–1923 (1999).

7. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B* **276**, 4261–4270 (2009).
8. Sempere, L. F., Martinez, P., Cole, C., Baguña, J. & Peterson, K. J. Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. *Evol. Dev.* **9**, 409–415 (2007).
9. Ruiz-Trillo, I. *et al.* A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. *Proc. Natl Acad. Sci. USA* **99**, 11246–11251 (2002).
10. Boursat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome structure of *Xenoturbella bocki* (phylum Xenoturbellida) is ancestral within the deuterostomes. *BMC Evol. Biol.* **9**, 107 (2009).
11. Philippe, H., Brinkmann, H., Martinez, P., Riutort, M. & Baguña, J. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS ONE* **2**, e717 (2007).
12. Rodríguez-Ezpeleta, N. *et al.* Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* **56**, 389–399 (2007).
13. Lartillot, N. & Philippe, H. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Phil. Trans. R. Soc. B* **363**, 1463–1472 (2008).
14. Ruiz Trillo, I., Riutort, M., Fourcade, H. M., Baguña, J. & Boore, J. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol. Phylog. Evol.* **33**, 321–332 (2004).
15. Papillon, D., Perez, Y., Caubit, X. & Le Parco, Y. Identification of chaetognaths as protostomes is supported by the analysis of their mitochondrial genome. *Mol. Biol. Evol.* **21**, 2122–2129 (2004).
16. Sperling, E. A. & Peterson, K. J. In *Animal Evolution. Genomes, Fossils and Trees* (eds Telford, M. J. & Littlewood, D. T. J.) Ch. 15, 157–170 (Oxford Univ. Press, 2009).
17. Lundin, K. Degenerating epidermal cells in *Xenoturbella bocki* (phylum uncertain). Nemertodermatida and Acoela (Platyhelminthes). *Belg. J. Zool.* **131**, 153–157 (2001).
18. Westblad, E. *Xenoturbella bocki* n.g., n.sp., a peculiar, primitive turbellarian type. *Arkiv Zool* **1**, 3–29 (1949).
19. Nielsen, C. After all: *Xenoturbella* is an acoelomorph! *Evol. Dev.* **12**, 241–243 (2010).
20. Franzen, A. & Afzelius, B. A. The ciliated epidermis of *Xenoturbella bocki* (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. *Zool. Scr.* **16**, 9–17 (1987).
21. Pardos, F. Fine structure and function of pharynx cilia in *Glossobalanus minutus* Kowalewsky (Entropneusta). *Acta Zool.* **69**, 1–12 (1988).
22. Tyler, S. In *Interrelationships of the Platyhelminthes* (eds Littlewood, D. T. J. & Bray, R. A.) 3–12 (Taylor & Francis, 2001).
23. Telford, M. J. *Xenoturbellida*: the fourth deuterostome phylum and the diet of worms. *Genesis* **46**, 580–586 (2008).
24. Baguña, J., Martinez, P., Paps, J. & Riutort, M. Back in time: a new systematic proposal for the Bilateria. *Proc. R. Soc. B* **363**, 1481–1491 (2008).
25. Hejnol, A. & Martindale, M. Q. M. Acoel development supports a simple planula-like urbilaterian. *Phil. Trans. R. Soc. B* **363**, 1493–1501 (2008).
26. Peterson, K. J., McPeck, M. A. & Evans, D. A. Tempo and mode of early animal evolution: inferences from rocks, Hox, and molecular clocks. *Paleobiology* **31**, 36–55 (2005).
27. Ruppert, E. E. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* **83**, 8–23 (2005).
28. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706–712 (2009).
29. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
30. Wheeler, B. *et al.* The deep evolution of metazoan microRNAs. *Evol. Dev.* **11**, 50–68 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Lartillot for reading the manuscript, W. Sterrer for helping collect material, and E. Sperling for help with small RNA library construction. H.P. is funded by Canada Research Chairs, Natural Sciences and Engineering Research Council and Réseau Québécois de Calcul de Haute Performance for computational resources: more than 220,000 central processing unit (CPU)-hours were used producing at least 7 tonnes of CO₂ excluding grey energy. R.R.C. and M.J.T. were part-funded by the Biotechnology and Biological Sciences Research Council SYNTAX scheme. K.J.P. is supported by the National Science Foundation and NASA Ames. R.R.C. was also supported by a Wellcome Trust core award, grant number 075491/Z/04. A.J.P. was supported by the Max-Planck Society for the Advancement of Sciences e.V. A.W. was funded by Inez Johanssons Stiftelse and Stiftelsen Lars Hiertas Minne.

Author Contributions H.P. and M.J.T. conceived and designed the study. M.J.T. assembled mitochondrial data. H.P. and H.B. assembled EST data and performed phylogenetic analyses of ESTs and mitochondrial genomes. M.J.T., L.L.M. and R.R.C. performed preliminary phylogenomic analyses. H.N. collected *Xenoturbella* for genomic and miRNA data. A.W. collected *Hofstenia*. K.J.P. and A.W. produced *Xenoturbella* and *Hofstenia* miRNA libraries. K.J.P. assembled and analysed the miRNA matrix. M.J.T., R.R.C. and A.J.P. produced *Xenoturbella* genomic data. M.J.T. drafted the paper with H.P. and K.J.P. All authors commented on the manuscript.

Author Information MicroRNA sequences are deposited in www.mirbase.org and can be found in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.T. (m.telford@ucl.ac.uk).

METHODS

Phylogenetic data assembly. Mitochondrial data. Metazoan mitochondrial protein coding genes were downloaded from OGRE (<http://drake.physics.mcmaster.ca/ogre/>). To assemble acoel partial genomes we used TBLASTN against EST collections from *Convolutiriba longifissura*, *Neochildia fusca* and *Symsagittifera roscoffensis* from the Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/>) at the National Center for Biotechnology Information. We used *Priapulius caudatus* mitochondrial proteins as a BLAST query. Open reading frames of positive hits were identified by aligning ESTs to the homologous protein sequence from *P. caudatus* using Genewise³¹. Multiple positives from a given species were then assembled into a contig using CAP3 (ref. 32). Nucleotides from each gene were aligned using TranslatorX³³ with the appropriate genetic code and using ClustalW³⁴ for the amino-acid alignment. Phylogenetic analyses were performed on the amino-acid translations.

Phylogenomic data. Alignments from Philippe *et al.*²⁸ and Dunn *et al.*³⁵ were updated with new sequences from GenBank (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein>), dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>) and the Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/>) at the National Center for Biotechnology Information. Single gene alignments were assembled using new features of the program Ed from the MUST software package³⁶. Ambiguously aligned regions were detected and removed with Gblocks³⁷ ($b2 = 75\%$, $b3 = 5$, $b4 = 5$, $b5 = \text{half}$); this automated selection was slightly refined by eye using Net (also from MUST).

Concatenations of single gene alignments into supermatrices were performed with scaFos³⁸. When multiple orthologous sequences were available for a particular operational taxonomic unit, scaFos helped to select the slowest-evolving sequence as determined from ML distances computed under a WAG + F model with TREE-PUZZLE³⁹. To minimize the amount of missing data, scaFos was allowed to create chimaerial operational taxonomic units by merging partial sequences from closely related species (Supplementary Table 1) when full-length sequences were not available. The amount of missing data per gene was limited to 25 species out of 66. Information about the names of the 197 selected genes, their size and the distribution of missing data are available in Supplementary Table 2.

To produce a data set that was tractable for the most time-consuming CAT + GTR + Γ model we reduced the number of taxa from 66 to 37. Our strategy for selecting taxa for elimination was as follows.

We first discarded seven species because of their incompleteness (that is, species with fewer than 16,000 amino acids); acoelomorphs, except the very incomplete *Convolutiriba longifissura* (5,458 amino acids), were exempt from this cull for obvious reasons.

Then we removed the most incomplete species within well-established clades. For example, within sponges, *Suberites* 23,000 amino acids versus 37,000 in *Amphimedon*; within urochordates, *Halocynthia* 24,000 amino acids versus 36,000 in *Molgula* and 37,000 in *Ciona*; within chelicerates, *Anoplodactylus* and *Acanthoscurria* 18,000 and 26,000 amino acids versus 37,000 in *Ixodes*.

This reduction of less complete taxa was balanced by the need to maintain a homogeneous taxon sampling (that is, about three species per major phylum Porifera, Cnidaria, Arthropoda, Mollusca, Annelida, Vertebrata). This strategy, while making the data set of a size that permits the use of the best (and most time-consuming) models, also allowed us to reduce the proportion of missing data from 30% (66 species) to 22% (37 species).

Another data set was assembled from the set of genes with the taxon sampling of Hejnol *et al.*⁷. In that case, only genes for which sequences were available for at least 55 species out of 94 were retained, yielding a set of 145 genes (24,632 unambiguously aligned positions, 30% missing data).

Phylogenetic inference. For mitochondrial and phylogenomic data sets, PhyloBayes analyses were performed with the CAT + Γ_4 mixture model. This accounts for across-site heterogeneities in the amino-acid replacement process²⁹. This model is implemented in an MCMC framework by the program PhyloBayes version 3.2 (ref. 40). Two independent runs were performed with a total length of 10,000 cycles (250 topological moves per cycle) with the same operators as in Lartillot *et al.*⁴¹, saved every ten cycles, for most data sets; however, for the super-matrix of 94 species, 20,000 cycles were necessary. The first 5,000 points were discarded as burn-in for all the data sets (except for the mitochondrial alignments where a burn-in of 1,000 was sufficient), and the posterior consensus was computed on the 500 (900 for mitochondrial alignments) remaining trees.

For the alignment of 66 species and 38,330 amino-acid positions, we applied a standard, time-consuming, bootstrap procedure⁴²: 100 pseudo-replicates were generated with SEQBOOT⁴³; each data set was analysed with PhyloBayes, trees were collected after the initial burn-in period and a consensus tree was computed by PhyloBayes; finally, a consensus tree was inferred from these 100 consensus trees using CONSENSE to compute the bootstrap support values for each node. To test the robustness of our results to gene sampling, we also performed a jack-knife

analysis of genes. We randomly sampled 50% of our 197 genes and for each of the ten replicates a PhyloBayes analysis was done using the CAT + Γ_4 model. The consensus trees obtained from all the post-burn-in trees (Supplementary Fig. 9) is identical to the tree based on the complete gene sample (Fig. 3), and jack-knife supports are very similar to bootstrap supports.

For the mitochondrial alignment (32 species and 2,118 positions) and a reduced EST-based alignment (37 species, 38,330 positions, 22% missing data), we also used the site-homogeneous GTR + Γ_4 and the time-consuming site-heterogeneous CAT + GTR + Γ_4 . We first performed statistical comparisons of the CAT + GTR + Γ_4 model, the CAT + Γ_4 model and the GTR + Γ_4 model using cross-validation tests as described in ref. 41. Ten replicates were run: 9/10 of the positions randomly drawn from the alignment were used as the learning set and the remaining 1/10 as the test set. For the GTR + Γ_4 model, MCMC were run for 1,100 cycles, 100 being discarded as burn-in. For the CAT + Γ_4 and CAT + GTR + Γ_4 models, MCMC were run for 1,600 (2,100) cycles with the mitochondrial (EST) alignment, 600 (1,100) being discarded as burn-in. Other matrix based models—WAG, JTT or LG, which are generally used—are special cases of the GTR model; for large data sets, the amount of data available is generally sufficient to learn the 190 free parameters of the GTR model⁴⁰; we verified by cross-validation that the GTR + Γ_4 model had a better fit to our data set than the WAG + Γ_4 model ($\Delta\ln L = 1219 \pm 60$). We therefore only used the GTR + Γ_4 model as the best site-homogeneous model in phylogenetic inference.

For these two smaller data sets (mitochondrial alignment of 32 species and 2,118 positions, and an EST-based alignment of 37 species and 38,330 positions), phylogenetic inference with the CAT + GTR + Γ_4 and GTR + Γ_4 models was performed with PhyloBayes 3.2 as for the CAT + Γ_4 model. For the GTR + Γ_4 model, the tree was also inferred with RAXML 7.0.4.1 (ref. 44), with 100 rapid bootstrap replicates, for the EST alignment.

For the six samples of the 37 species data set, we inferred the trees with the GTR + Γ_4 model using RAXML and PhyloBayes. We expect minor, and even no, differences between the ML and Bayesian inference because the same model is used (priors are known to have an effect on Bayesian analysis, but it should be very small given the large size of the data set).

In five of these taxon samples, the RAXML and PhyloBayes topologies are identical, only the position of *Trichoplax* varies in some cases, but bootstrap supports for the placement of *Trichoplax* are around 50%. In one taxon sample (that excluding nemertodermatids), various chains of PhyloBayes failed to converge towards the same topology, differing only by the position of Acoela (which corresponds to bootstrap support close to 30%); interestingly, all the topologies found by PhyloBayes were also found among the bootstrap replicates of RAXML (corresponding to bootstrap support between 10 and 20%). This indicates that the various topologies have very similar likelihoods and that PhyloBayes is unable to switch readily among these various local minima, at least in a reasonable time (that is, several months of computation).

Compositional heterogeneity. The amino-acid composition of the 66 species EST-based data set was visualized by assembling a 20×66 matrix containing the frequency of each amino acid per species using the program Net from the MUST package³⁶. This matrix was then displayed as a two-dimensional plot in a principal component analysis, as implemented in the R package. Supplementary Fig. 7 demonstrates that the amino-acid compositions of *Xenoturbella*, Nemertodermatida and Acoela are not similar and therefore that their monophyly is not likely to be due to a compositional artefact.

To verify that amino-acid compositional heterogeneity does not bias our inference, we cannot use the time-heterogeneous CAT-BP⁴⁵ because of the intractable computational burden. Instead, we used the Dayhoff coding⁴⁶ in which the amino acids are recoded according to the six classes defined by M. Dayhoff. The recoded alignments were analysed with PhyloBayes 3.2 using the CAT + Γ_4 model, under the same conditions as previously. The resulting phylogeny (Supplementary Fig. 8) is almost identical to the tree of Fig. 3; some minor rearrangements within Porifera, Lophotrochozoa, Vertebrata and Xenacoelomorpha correspond to poorly supported groups. Our inference thus does not seem to be biased by compositional heterogeneity.

The mitochondrial analysis is complicated to an unknown extent by the existence of multiple variants of the genetic code within deuterostomes. The resulting compositional biases may impede correct reconstruction of relationships within the deuterostomes¹. Different mitochondrial genetic codes are found in vertebrates, urochordates, cephalochordates, echinoderms and hemichordates, and the observation that the acoelomorph and *Xenoturbella* code (invertebrate code) differs from all of these makes the deuterostome affinity observed in our analyses conservative.

The non-monophyly of Cnidaria observed in the mitochondrial tree is likely to be incorrect. This problem is particularly difficult because of the extreme rate heterogeneity in the tree (the distance between Porifera and Anthozoa is smaller

than the distances within Echinodermata). This heterogeneity is coupled with a change in the properties of the evolutionary process⁴⁷. Importantly, we do not see any tree reconstruction artefact that would erroneously cluster the fast-evolving acoels with the slow-evolving *Xenoturbella* (whereas the fast rate of *Aurelia* easily explains its incorrect position by an attraction with the very-long-branched Bilateria). As a result, it is reasonable to attribute the position of Acoela to genuine phylogenetic signal rather than to non-phylogenetic signal.

miRNA data collection. Specimens of *Xenoturbella bocki* were collected as previously described² and were starved for 5 months to avoid contamination by their food. Specimens of *H. miamia* were extracted from algae and leaf litter collected among mangroves at Walsingham Pond, Bermuda. The worms were starved for 2 weeks before miRNA extraction. Small RNA libraries were constructed and sequences analysed as described elsewhere³⁰. miRNA presequences were also recovered from *Xenoturbella* genomic DNA traces by BLAST searches.

31. Birney, E., Clamp, M. & Durbin, R. GeneWise and GenomeWise. *Genome Res.* **14**, 988–995 (2004).
32. Huang, X. & Madan, A. CAP3: a DNA assembly programme. *Genome Res.* **9**, 868–877 (1999).
33. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
34. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405 (1998).
35. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
36. Philippe, H. MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res.* **21**, 5264–5272 (1993).
37. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
38. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7** (Suppl. 1), S2 (2007).
39. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504 (2002).
40. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
41. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7** (Suppl. 1), S4 (2007).
42. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
43. Felsenstein, J. *PHYMLIP (Phylogeny Inference Package)* version 3.69 (Department of Genome Sciences, Univ. Washington, Seattle, 2005).
44. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
45. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **25**, 842–858 (2008).
46. Hrdy, I. *et al.* *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**, 618–622 (2004).
47. Roure, B. & Philippe, H. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evol. Biol.* (in the press).

Transmembrane semaphorin signalling controls laminar stratification in the mammalian retina

Ryota L. Matsuoka^{1,3}, Kim T. Nguyen-Ba-Charvet^{4,5,6}, Aijaz Parray^{4,5,6}, Tudor C. Badea^{2,3,†}, Alain Chédotal^{4,5,6} & Alex L. Kolodkin^{1,3}

In the vertebrate retina, establishment of precise synaptic connections among distinct retinal neuron cell types is critical for processing visual information and for accurate visual perception. Retinal ganglion cells (RGCs), amacrine cells and bipolar cells establish stereotypic neurite arborization patterns to form functional neural circuits in the inner plexiform layer (IPL)^{1–3}, a laminar region that is conventionally divided into five major parallel sublaminae^{1,2}. However, the molecular mechanisms governing distinct retinal subtype targeting to specific sublaminae within the IPL remain to be elucidated. Here we show that the transmembrane semaphorin *Sema6A* signals through its receptor *PlexinA4* (*PlexA4*) to control lamina-specific neuronal stratification in the mouse retina. Expression analyses demonstrate that *Sema6A* and *PlexA4* proteins are expressed in a complementary fashion in the developing retina: *Sema6A* in most ON sublaminae and *PlexA4* in OFF sublaminae of the IPL. Mice with null mutations in *PlexA4* or *Sema6A* exhibit severe defects in stereotypic lamina-specific neurite arborization of tyrosine hydroxylase (TH)-expressing dopaminergic amacrine cells, intrinsically photosensitive RGCs (ipRGCs) and calbindin-positive cells in the IPL. *Sema6A* and *PlexA4* genetically interact *in vivo* for the regulation of dopaminergic amacrine cell laminar targeting. Therefore, neuronal targeting to subdivisions of the IPL in the mammalian retina is directed by repulsive transmembrane guidance cues present on neuronal processes.

Synaptic connections among distinct neuronal cell types are organized in specific laminae within many regions of the nervous system. In the vertebrate retina, RGCs, amacrine cells, and bipolar cells have multiple morphologically distinct subtypes (RGCs, approximately 20; amacrine cells, approximately 30; bipolar cells, approximately 12), and each subtype elaborates a characteristic sublaminal connection pattern within the IPL^{1,3}. Recent studies have shown that homophilic cell adhesion molecules, including sidekicks and Dscams, direct sublaminal targeting of distinct amacrine, bipolar and RGC cell types in the developing chicken retina^{4,5}. A mutation in mouse *Dscam* disturbs process self avoidance, mosaic spacing and stratification of several amacrine cell subtypes^{6,7}; however, it is not clear whether *Dscam* regulates the stratification of these amacrine cell subtypes directly or whether this is a consequence of other abnormalities in the *Dscam* mutant mouse retina, including disorganization of retinal layers and an expanded IPL. Thus, molecular cues that organize specific laminar stratifications in the mammalian retina have yet to be defined.

The semaphorin family of guidance cues includes secreted and membrane-bound proteins that have key roles in various neuronal developmental processes, including axon guidance and branching, neuronal migration and dendritic arborization⁸. Multiple classes of semaphorins have been shown to be expressed in the developing mammalian retina^{9,10}; however, whether or how semaphorins function within the retina is not known.

To assess the *in vivo* roles of semaphorins and their receptors in retinal development, we first conducted expression analyses for conventional semaphorin receptors, *neuropilins* (*Npn-1* and *Npn-2*) and *plexins* (*PlexA1–A4*, *B1–B3*, *C1*, *D1*) in the developing mouse retina by *in situ* hybridization. We observed that multiple *plexins* and *neuropilins* are expressed both in overlapping and in distinct locations in the developing retina (data not shown). To investigate physiological functions of these semaphorin receptors in retinal development, we analysed mice harbouring targeted mutations in genes encoding each plexin and neuropilin by immunohistochemistry using various retinal markers, including Pax6, Chx10, Thy-1, TH, calbindin, choline acetyltransferase (ChAT), calretinin and protein kinase C alpha (PKC- α) (Supplementary Fig. 1 and data not shown)^{11–13}. We identified defects in the stereotypic lamina-specific neurite arborization of tyrosine hydroxylase-positive (TH⁺) dopaminergic amacrine cells, and calbindin-positive cells, in the IPL of adult mice homozygous for a targeted mutation in the gene encoding the *PlexA4* receptor¹¹ (Fig. 1b, d). We used division of the IPL into five parallel sublaminae (S1–S5, S5 being closest to the ganglion cell layer) for our analyses, as previously described^{1,2}. We observed that dopaminergic amacrine cells, which predominantly stratify in the S1 sublamina of the IPL in wild-type retinas (Fig. 1a and Supplementary Fig. 2a, c), extend aberrant processes into S4/S5 in the *PlexA4*^{−/−} mutant retina (Fig. 1b and Supplementary Fig. 2b, d). Similarly, calbindin-positive cells, which typically establish their projections in three strata at the borders of S1 and S2, S2 and S3, S3 and S4 in the IPL of wild-type retinas¹ (Fig. 1c), showed aberrant targeting of their processes to S4/S5 in the *PlexA4*^{−/−} retina (Fig. 1d). These sublaminal targeting defects observed in dopaminergic amacrine cells and calbindin-positive cells show full penetrance and expressivity in *PlexA4*^{−/−} mutant retinas (*n* = 10 mutant animals). To determine precisely where the aberrant processes of dopaminergic amacrine cells and calbindin-positive cells are localized within the *PlexA4*^{−/−} retina, we performed double-immunolabelling to visualize these two neuronal subtypes and also cells labelled with an antibody directed against calretinin, which marks three strata at the borders of S1 and S2, S2 and S3, S3 and S4 (ref. 1) (the localization of calretinin⁺ processes is not disrupted in *PlexA4*^{−/−} retinas; Fig. 1e–f). We found that the aberrant processes of both TH⁺ and calbindin⁺ neuronal subtypes were localized predominantly adjacent to the calretinin⁺ S3/S4 stratification band within S4/S5 in *PlexA4*^{−/−} retinas (Fig. 1g, h). We confirmed that these two populations of mistargeted retinal neurons labelled by anti-TH and anti-calbindin are different retinal subtypes (Supplementary Fig. 3), demonstrating that *PlexA4* directs distinct retinal subtype targeting in the IPL *in vivo*. We also found that the calbindin⁺ cells exhibiting neurite arborization defects in *PlexA4*^{−/−} retinas are most probably amacrine cells because these calbindin⁺ cells with aberrant processes in the S4/S5 sublaminae are co-immunolabelled by syntaxin, a pan-amacrine cell marker, but not by Brn3a, a marker for a subset of RGCs (data not shown).

¹The Solomon H. Snyder Department of Neuroscience, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ²Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ³Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁴Institut National de la Santé et de la Recherche Médicale (INSERM), UMR S968, Institut de la Vision, F-75012 Paris, France. ⁵Université Pierre et Marie Curie (UPMC) Paris VI, UMR S968, Institut de la Vision, F-75012 Paris, France. ⁶Centre National de la Recherche Scientifique (CNRS) UMR 7210, Institut de la Vision, F-75012 Paris, France. [†]Present address: Retinal Circuit Development & Genetics Unit, Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute, Bethesda, Maryland 20892, USA.

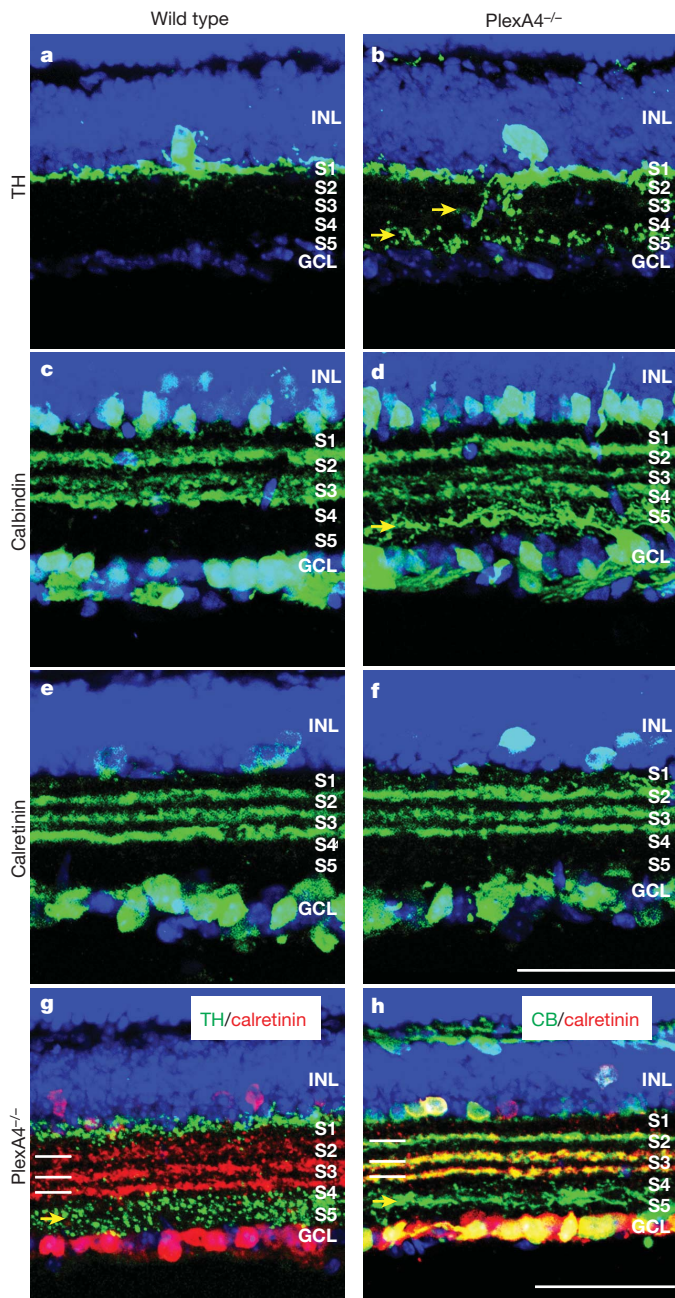


Figure 1 | PlexinA4 directs lamina-specific neurite arborization of dopaminergic amacrine cells and calbindin-positive cells in the IPL *in vivo*. **a–f**, Wild-type (**a**, **c**, **e**) and *PlexA4*^{−/−} (**b**, **d**, **f**) adult retina sections were immunostained with antibodies against TH (**a**, **b**), calbindin (**c**, **d**) and calretinin (**e**, **f**). INL, inner nuclear layer; GCL, ganglion cell layer. In *PlexA4*^{−/−} retinas, TH-positive dopaminergic amacrine cells and calbindin-positive cells exhibit defects in lamina-specific neurite arborization (yellow arrows in **b** and **d**, *n* = 10 *PlexA4*^{−/−} animals). In wild-type retinas, dopaminergic amacrine cell processes are observed predominantly in the S1 sublamina of the IPL (**a**). In contrast, aberrant punctate immunostaining is detected in the S4/S5 sublaminae, in addition to S1, in all *PlexA4*^{−/−} retinas examined (**b**). The normal stratification of calbindin-positive cells in the IPL (**c**) is disrupted in *PlexA4*^{−/−} retinas, resulting in aberrant processes in S4/S5 (**d**). Calretinin-positive cells show normal sublaminal stratification in the IPL of *PlexA4*^{−/−} retinas (**e**, **f**). **g**, **h**, *PlexA4*^{−/−} adult retina sections double-immunostained with anti-calretinin (white bars) and anti-TH (**g**), or anti-calretinin (white bars) and anti-calbindin (CB) (**h**). Aberrant processes in *PlexA4*^{−/−} retinas from dopaminergic amacrine cells, and from calbindin-positive cells, are found closer to the GCL than the calretinin-positive processes that lie between S3 and S4 in the IPL (yellow arrows). Scale bars, 50 μ m in **h** for **a**, **b**, **g**, **h**, and in **f** for **c–f**.

In contrast, other subtypes of RGCs and amacrine cells, including AII amacrine cells labelled with Disabled-1 (Dab-1), vGlut3-positive amacrine cells, cholinergic amacrine cells labelled with ChAT and R-cadherin-positive cells, show normal neurite arborization in the IPL of *PlexA4*^{−/−} mutant retinas (Supplementary Fig. 4). This result further demonstrates that PlexA4 regulation of lamina-specific neurite arborization of retinal neuronal subtypes in the IPL is cell-type specific. Dopaminergic amacrine cell processes, which normally are targeted exclusively to the S1 sublamina (the OFF layer), are misguided in *PlexA4*^{−/−} mutants to the S4/S5 sublaminae (a portion of the ON layer), suggesting that PlexA4 contributes to the segregation of ON and OFF layers within the IPL¹⁴.

Recent studies have shown that dopaminergic amacrine cells co-stratify with, and are synaptically coupled to, M1-type melanopsin intrinsically photosensitive retinal ganglion cells (ipRGCs) in S1 of the mouse IPL^{15,16}. Therefore, we asked whether the abnormality in dopaminergic amacrine cell process stratification affects M1-type ipRGC dendritic arborization in the *PlexA4*^{−/−} retina. We used an antibody directed against the carboxy (C) terminus of rat melanopsin to label M1-type melanopsin ipRGCs^{17,18}, and we observed aberrant dendritic arborization of M1-type ipRGCs in S4/S5, in addition to stratification within S1 of the IPL in *PlexA4*^{−/−} retina (Fig. 2a top, b top). We also observed that approximately 75% of aberrant dopaminergic amacrine cell processes (TH-immunoreactive puncta) were co-localized with M1-type ipRGC dendrites (C-terminal melanopsin-immunoreactive puncta) within S4/S5 in *PlexA4*^{−/−} mutant retinas (Fig. 2a bottom, b bottom, c, d), suggesting that synaptic connectivity between dopaminergic amacrine cells and M1-type ipRGCs may be still preserved, even though these two neuronal populations have mispositioned processes within the IPL of the *PlexA4*^{−/−} mutant retina. Therefore, PlexA4 is required for precise sublaminal targeting of dopaminergic amacrine cell processes and M1-type ipRGC dendrites within the IPL but may not be essential for synaptic target selection between these two neuronal populations. We examined both the cell number and mosaic patterning of dopaminergic amacrine cells and ipRGCs; we observed no significant difference between wild-type and *PlexA4*^{−/−} retinas (Supplementary Fig. 5). We also observed no evidence of neuronal process self-avoidance deficits in these cell types in *PlexA4*^{−/−} retinas (Supplementary Fig. 5a–b, e–f). In addition, *PlexA4*^{−/−} RGC axons do not exhibit major projection defects in their trajectories to image-forming and non-image-forming targets within the brain (Supplementary Fig. 6); nor are errors observed in bipolar cell axon targeting within the IPL (Supplementary Fig. 7).

We next analysed PlexA4 protein expression using a PlexA4-specific antibody¹⁹. We observed no immunostaining with this antibody in *PlexA4*^{−/−} retinas, confirming its specificity (Supplementary Fig. 8a, b). Anti-PlexA4 immunostaining at different postnatal ages shows that PlexA4 is strongly expressed on neuronal processes that predominantly stratify in S1 and S2 of the developing IPL (Fig. 3a left, b). We also observed that dopaminergic amacrine cell processes in S1 co-stratified with the upper PlexA4⁺ S1 band (Fig. 3d left, middle and right). To test further if dopaminergic amacrine cells express PlexA4, we performed *in situ* hybridization experiments using a *PlexA4* antisense probe followed by anti-TH immunolabelling. We found that *PlexA4* messenger RNA (mRNA) is localized to the cell bodies of dopaminergic amacrine cells (26 out of 26 dopaminergic amacrine cells analysed showed co-localization of TH and *PlexA4* mRNA; Fig. 3e left, middle and right). Taken together, these results strongly suggest PlexA4 functions cell-autonomously in dopaminergic amacrine cells to regulate stratification of this cell type within the IPL. We did not observe PlexA4 protein expression in M1-type ipRGC cell bodies and dendrites (Fig. 3f left and right). This suggests that the M1-type ipRGC dendritic stratification deficit within the IPL of *PlexA4*^{−/−} retinas is a secondary consequence of the developmental defects observed in the IPL of *PlexA4*^{−/−} retinas, supporting a primary role for amacrine cells in directing RGC dendritic stratification^{3,20,21}. Given our observation that

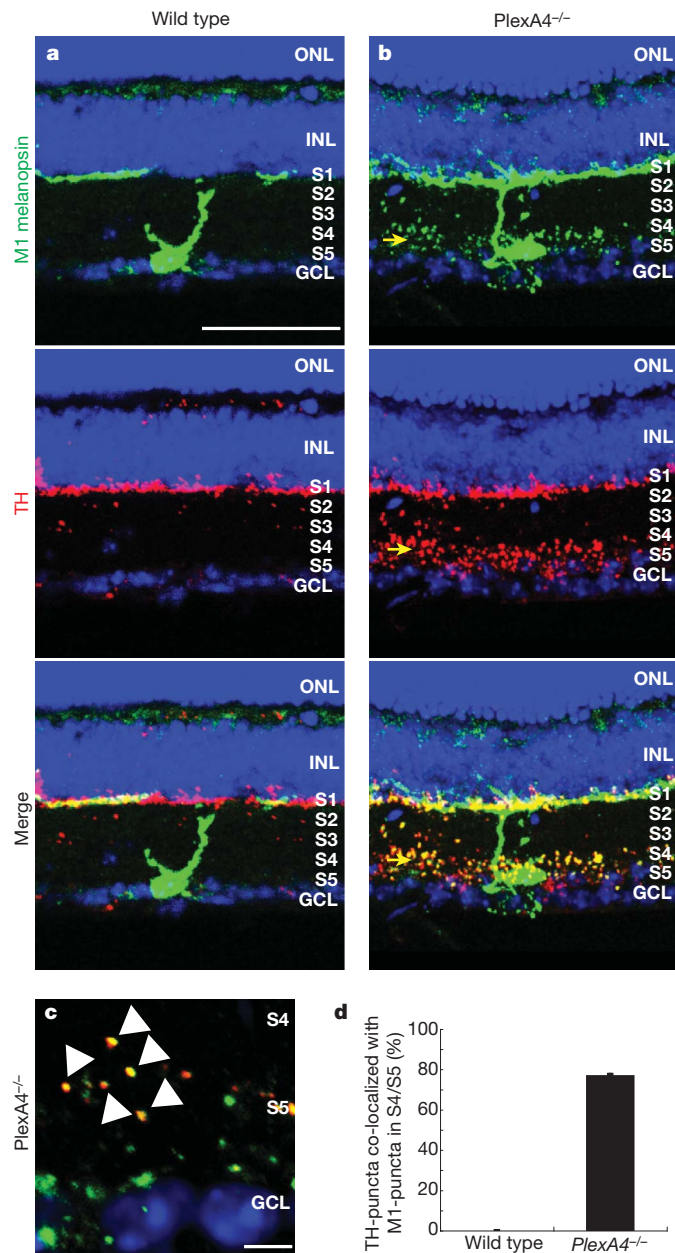


Figure 2 | PlexinA4 controls dendritic targeting of M1-type ipRGCs within the IPL, but not co-localization of dopaminergic amacrine cell and ipRGC processes. **a, b**, Top, middle, bottom, double-immunostaining using antibodies directed against the C terminus of rat melanopsin (**a, b**, top, green) and against TH (**a, b**, middle, red) of wild-type (**a**, top, middle, bottom) and *PlexA4*^{-/-} (**b**, top, middle, bottom) adult retina sections (merged in **a, b**, bottom). Ectopic dendritic processes of M1-type ipRGCs were observed in the S4/S5 sublaminae of *PlexA4*^{-/-} retinas, as were aberrant dopaminergic amacrine processes (yellow arrows, **a**, top, middle, bottom, $n = 4$ mutant animals). Wild-type M1-type ipRGC dendritic processes and dopaminergic amacrine cell processes are only observed in S1 (**a**, top, middle, bottom). **c**, High-magnification view of S4/S5 in *PlexA4*^{-/-} retinas double-immunostained with anti-C-terminal melanopsin and anti-TH. Most TH-positive puncta are co-localized with melanopsin-positive puncta (white arrowheads). **d**, Quantification of ectopic TH-positive puncta co-localized with the ectopic M1-type melanopsin puncta in S4/S5 of *PlexA4*^{-/-} retinas. Nearly 76% (194 TH-positive puncta among a total of 254 puncta) of the ectopic TH-positive puncta were co-localized with ectopic M1-type melanopsin puncta in S4/S5 ($76.4 \pm 1.2\%$ co-localization). In wild-type retinas, almost no TH-positive puncta were observed in S4/S5. Error bar, s.e.m. ($n = 3$ animals per genotype). Scale bars, 50 μm in **a, b**, top, middle, bottom, 5 μm in **c**.

dopaminergic amacrine cell processes and M1-type ipRGC dendrites are co-localized in the S4/S5 sublaminae of *PlexA4*^{-/-} retinas, dopaminergic amacrine cells may provide specific cues used by M1-type ipRGCs to form selective synaptic contacts.

There are two major classes of potential PlexA4 ligand: secreted class 3 semaphorins that bind to neuropilin-obligate co-receptors and form a holoreceptor complex with PlexA4 (ref. 8); and transmembrane class 6 semaphorins that directly bind to PlexA4 in a neuropilin-independent manner⁸. We first analysed neurite arborization of dopaminergic amacrine cells, M1-type ipRGCs and calbindin-positive cells in *Npn-1*^{Sema-/-Sema-} (an *Npn-1* allele that generates a variant Npn-1 protein incapable of responding to class 3 semaphorin signalling) and *Npn-2*^{-/-} mutant retinas^{12,13}. We observed normal neurite stratification patterns in the IPL of all of these three neuronal subtypes in both *Npn-1*^{Sema-/-Sema-} and *Npn-2*^{-/-} retinas (Supplementary Fig. 9), indicating that secreted class 3 semaphorins are unlikely to act as ligands for PlexA4 in the retina.

Transmembrane class 6 semaphorins, including Sema6A, bind directly to PlexA4¹⁹. Sema6A induces growth cone collapse of several neuronal subtypes through the PlexA4 receptor *in vitro* and acts as a repulsive ligand for PlexA4 *in vivo*, regulating hippocampal mossy fibre projections and corticospinal tract decussation^{19,22}. To ask whether Sema6A is a PlexA4 ligand, required for normal retinal development, we first analysed Sema6A protein expression using a Sema6A-specific antibody²³ (Supplementary Fig. 8c, c', d, d'). We found that Sema6A protein is strongly expressed in retinal S3b–S5 sublaminae (S3b being approximately the lower half of S3), and expressed at much lower levels in the S1–S3a sublaminae (S3a being approximately the upper half of S3) (Fig. 3c). We double-immunolabelled retinal sections with Sema6A and PlexA4 antibodies and found that strong Sema6A and PlexinA4 protein immunoreactivity is detected in adjacent regions of the developing IPL throughout early postnatal retinal development (Fig. 3a left, middle and right and Supplementary Fig. 10). These results support the hypothesis that Sema6A functions as a repulsive barrier within the developing IPL for neuronal processes expressing PlexA4, including dopaminergic amacrine cells. We observed Sema6A is not expressed in dopaminergic amacrine cells or M1-type ipRGCs (Fig. 3g, h), consistent with Sema6A serving a non-cell autonomous role in constraining the targeting of processes from these neuronal cell types in the IPL. However, immunolabelling experiments revealed that RGC and amacrine cell subtypes, distinct from ipRGCs and dopaminergic amacrine cells, are the major cellular sources of Sema6A protein in the developing IPL (see Supplementary Figs 11–13).

Phenotypic analysis of mice homozygous for a targeted gene-trap mutation in the *Sema6A* locus⁹ showed that *Sema6A*^{-/-} mutants phenocopy the neurite stratification defects in dopaminergic amacrine cells, M1-type ipRGCs and calbindin-positive cells we observe in *PlexA4*^{-/-} mutant retinas (with full penetrance and expressivity, $n = 8$ *Sema6A*^{-/-} mutant animals, Fig. 4a–f). This result strongly suggests that Sema6A is a functional ligand for PlexA4, required for regulating select aspects of retinal neurite stratification *in vivo*. To assess further the ligand–receptor relationship between Sema6A and PlexA4 in retinal development *in vivo*, we investigated genetic interactions between *PlexA4* and *Sema6A* by analysing mice doubly heterozygous for *Sema6A* and *PlexA4* mutations. We quantified the number of TH-positive immunoreactive puncta localized in S4/S5 of *Sema6A*^{+/-};*PlexA4*^{+/-} mutant mice (Fig. 4k). In wild-type retinas, TH⁺ immunoreactive puncta in S4/S5 were almost undetectable (Fig. 4g). Mice heterozygous for either *PlexA4* or *Sema6A* mutations did not show a significant increase in the number of the TH⁺ puncta in S4/S5 (Fig. 4h, i). However, *Sema6A*^{+/-};*PlexA4*^{+/-} mutant mice exhibited a markedly increased number of the TH⁺ puncta in S4/S5 (Fig. 4j). Therefore, *Sema6A* and *PlexA4* functionally interact *in vivo* and probably act in a common signalling pathway. Together with the complementary expression patterns of Sema6A and PlexA4 in specific regions of the developing IPL, these results strongly support a model in

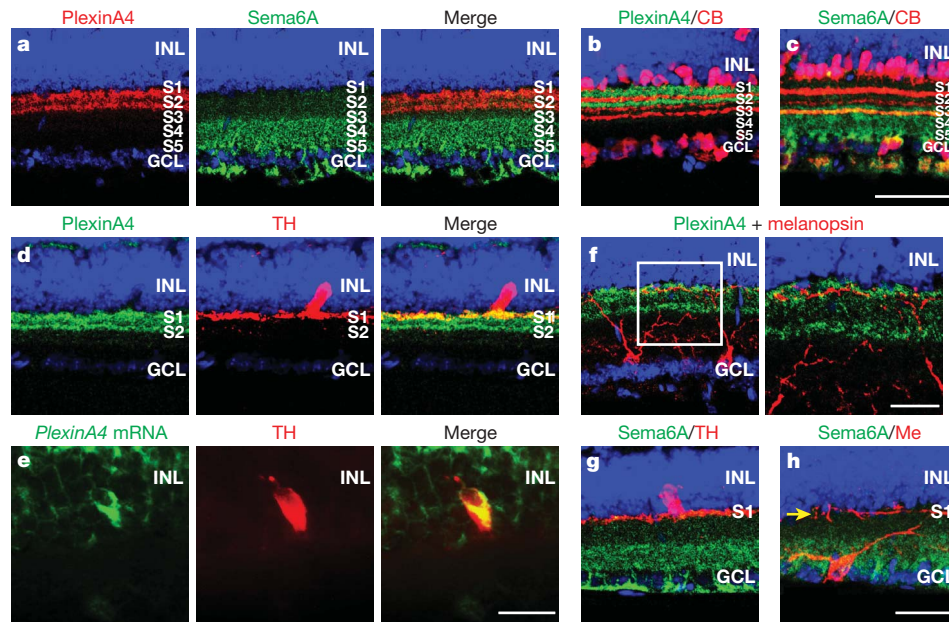


Figure 3 | PlexinA4 and Sema6A exhibit complementary protein expression in the developing mouse retina. **a**, Left, middle and right, P14 retina section double-immunostained with anti-PlexA4 (**a**, left, red) and anti-Sema6A (**a**, middle, green) (merged in **a**, right). Strong Sema6A immunoreactivity in the IPL was observed in approximately one-half of S3 and throughout S4 and S5, whereas PlexA4 expression is stratified in two distinct layers in S1 and S2. **b**, P14 retina section double-immunostained with anti-PlexA4 (green) and anti-calbindin (CB, red) shows PlexA4 protein localization in S1/S2 sublaminae relative to calbindin-positive neuronal processes. **c**, P14 retina section double-immunostained with anti-Sema6A (green) and anti-calbindin (CB, red) shows Sema6A protein localization in S3–S5 relative to calbindin-positive neuronal processes. **d**, Left, middle and right, P14 retina section double-immunostained with anti-PlexA4 (**d**, left, green) and anti-TH (**d**, middle, red), revealing colocalization of PlexA4 and TH immunoreactivity in S1 of the IPL (merged in **d**, right). **e**, Left, middle, right, P14 retina section hybridized with *PlexA4*

antisense probe (**e**, left, green) followed by anti-TH immunolabelling (**e**, middle, red, merged in **e**, right). *PlexA4* mRNA is localized to the cell body of dopaminergic amacrine cells (of 26 TH-positive amacrine cells scored, all were positive for *PlexA4* mRNA). **f**, Left and right, P14 retina section double-immunostained with anti-PlexA4 (green) and anti-amino (N)-terminal melanopsin (red), which labels multiple ipRGC subtypes¹⁸ (**f**, left; high magnification of the area in the white square shown in **f**, right). PlexA4 immunoreactivity was not observed in the cell bodies or dendrites of ipRGCs. **g**, **h**, P14 retina sections double-immunostained with anti-Sema6A (green) and anti-TH (**g**, red) or anti-N-terminal melanopsin (Me) (**h**, red). Sema6A protein was not observed in cell bodies, or processes, of dopaminergic amacrine cells and ipRGCs (M1-type ipRGC dendritic processes in the S1 indicated by yellow arrow in **h**). Scale bars, 50 μm in **c** for **a**, left, middle and right, **b**, **c**, 50 μm in **h** for **d**, left, middle and right, **f**, left, **g**, **h**, 20 μm in **e**, right for **e**, left, middle and right, 20 μm in **f**, right.

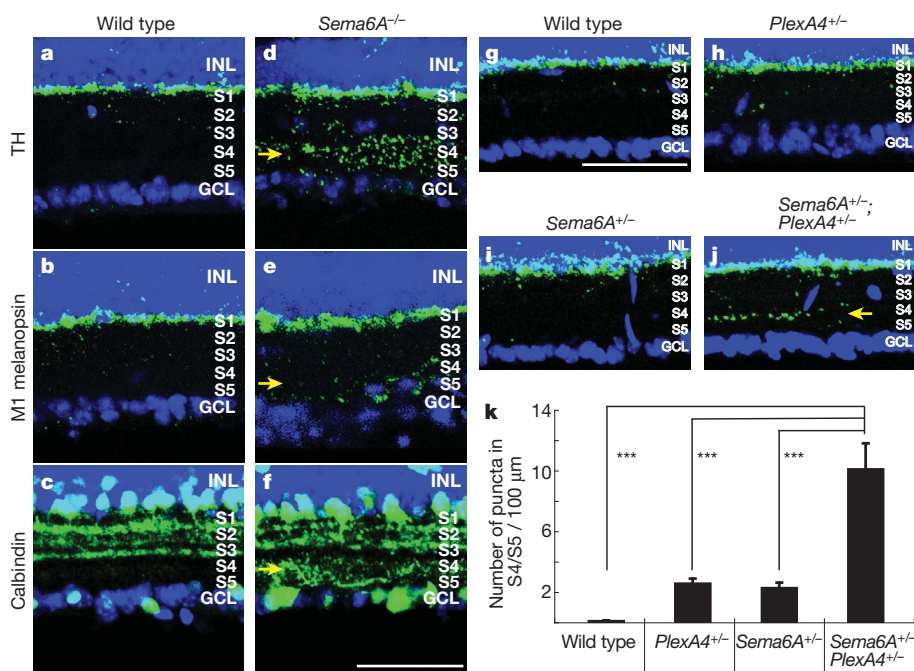


Figure 4 | Sema6A signalling through the PlexinA4 receptor directs retinal sublamina targeting. **a–f**, Wild-type (**a–c**) and *Sema6A*^{−/−} (**d–f**) adult retina sections were immunostained with antibodies against TH (**a**, **d**), C-terminal melanopsin (**b**, **e**) and calbindin (**c**, **f**). *Sema6A*^{−/−} retinas recapitulate the lamina-specific neurite arborization defects of dopaminergic amacrine cells, M1-type ipRGCs and calbindin-positive cells (yellow arrows) observed in *PlexA4*^{−/−} retinas ($n = 8$ *Sema6A*^{−/−} animals). **g–j**, Wild-type (**g**), *PlexA4*^{+/-} (**h**), *Sema6A*^{+/-} (**i**) and *Sema6A*^{+/-}; *PlexA4*^{+/-} (**j**) adult retina sections were immunostained with anti-TH. **k**, Quantification of ectopic TH-positive puncta detected in the S4/S5 sublaminae in wild-type (**g**), *PlexA4*^{+/-} (**h**), *Sema6A*^{+/-} (**i**) and *Sema6A*^{+/-}; *PlexA4*^{+/-} (**j**) sections ($n = 4$ animals for each genotype). An increased number of TH-positive puncta were observed in S4/S5 in *Sema6A*^{+/-}; *PlexA4*^{+/-} retinas (10.1 ± 1.7 puncta per 100 μm, yellow arrow, **j**) compared with the other three genotypes (0.1 ± 0.1 puncta per 100 μm, wild type; 2.6 ± 0.3 puncta per 100 μm, *PlexA4*^{+/-}; 2.3 ± 0.4 puncta per 100 μm, *Sema6A*^{+/-}; **g–i**). Error bars, s.e.m. *** $P < 0.01$ by Tukey's multiple comparison test. Scale bars, 50 μm in **f** for **a–f**, and in **g** for **g–j**.

which *Sema6A* acts as a ligand for the *PlexA4* receptor to regulate dopaminergic amacrine cell process targeting in the IPL.

We provide here a demonstration of a molecular cue that directs lamina-specific neurite arborization in the developing mouse retina. We show that *Sema6A* and its receptor *PlexA4* exhibit complementary expression patterns throughout postnatal IPL development, that *Sema6A*^{-/-} mutant mice phenocopy defects in lamina-specific neurite stratification of specific retinal neuron subtypes observed in *PlexA4*^{-/-} mutant mice, and that they functionally interact *in vivo*. *PlexA4*^{-/-} mutant retinas do not exhibit defects in neurite fasciculation of the retinal cell types that show defects in sublamina targeting in the IPL (Supplementary Fig. 5), further suggesting that sublamina targeting in the vertical plane of the mouse retina and neurite arborization in the horizontal plane of the mouse retina are governed by separate mechanisms. Our observations of *Sema6A* and *PlexA4* function in retinal development suggest that initial laminar targeting to broad regions within the developing mouse IPL is directed by a transmembrane guidance cue located on neuronal processes that signals through its receptor, present on other neuronal subtypes. This defines a heterophilic interaction distinct from homophilic adhesive interactions mediated by molecules such as sidekicks and Dscams. Neuronal circuitry mediating two parallel ON/OFF visual pathways is spatially segregated in the IPL of the vertebrate retina^{1,14}, and this spatial segregation has crucial roles in the effective transmission of distinct light responses to the brain¹. Determining how the ON and OFF pathways are segregated at the circuit level is fundamental for understanding visual perception: our results suggest that these distinct neuronal pathways are established in the IPL through the action of a transmembrane guidance cue and its receptor. Our elucidation of molecular events critical for lamina-specific targeting in the IPL of the mammalian retina may have general implications for understanding mechanisms that govern the establishment of neuronal connectivity, in particular how laminar organization is achieved during neural development.

METHODS SUMMARY

The day of birth in this study is designated as postnatal (P) day 0. The *PlexA4*-deficient and *Sema6A* gene-trap mouse lines were previously described^{9,11}. Immunohistochemistry and *in situ* hybridization were performed as previously described¹³, and the retinal regions we imaged did not include areas near the peripheral edges or the optic nerve head of retinas. See Methods for additional detailed experimental procedures, including wholemount retina staining, density recovery profile analysis, quantification of anti-TH and anti-C-terminal melanopsin co-localized puncta, genetic interaction analysis, X-gal staining, cholera toxin injection and statistical analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 September; accepted 17 November 2010.

Published online 26 January 2011; corrected 10 February 2011 (see full-text HTML version for details).

- Wassle, H. Parallel processing in the mammalian retina. *Nature Rev. Neurosci.* **5**, 747–757 (2004).
- Sanes, J. R. & Zipursky, S. L. Design principles of insect and vertebrate visual systems. *Neuron* **66**, 15–36 (2010).
- Huberman, A. D., Clandinin, T. R. & Baier, H. Molecular and cellular mechanisms of lamina-specific axon targeting. *Cold Spring Harb Perspect Biol* **2**, a001743 (2010).
- Yamagata, M., Weiner, J. A. & Sanes, J. R. Sidekicks: synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell* **110**, 649–660 (2002).

- Yamagata, M. & Sanes, J. R. Dscam and Sidekick proteins direct lamina-specific synaptic connections in vertebrate retina. *Nature* **451**, 465–469 (2008).
- Fuerst, P. G., Koizumi, A., Masland, R. H. & Burgess, R. W. Neurite arborization and mosaic spacing in the mouse retina require DSCAM. *Nature* **451**, 470–474 (2008).
- Fuerst, P. G., Harris, B. S., Johnson, K. R. & Burgess, R. W. A novel null allele of mouse DSCAM survives to adulthood on an inbred C3H background with reduced phenotypic variability. *Genesis* **48**, 578–584 (2010).
- Tran, T. S., Kolodkin, A. L. & Bharadwaj, R. Semaphorin regulation of cellular morphology. *Annu. Rev. Cell Dev. Biol.* **23**, 263–292 (2007).
- Leighton, P. A. *et al.* Defining brain wiring patterns and mechanisms through gene trapping in mice. *Nature* **410**, 174–179 (2001).
- de Winter, F., Cui, Q., Symons, N., Verhaagen, J. & Harvey, A. R. Expression of class-3 semaphorins and their receptors in the neonatal and adult rat retina. *Invest. Ophthalmol. Vis. Sci.* **45**, 4554–4562 (2004).
- Yaron, A., Huang, P. H., Cheng, H. J. & Tessier-Lavigne, M. Differential requirement for Plexin-A3 and -A4 in mediating responses of sensory and sympathetic neurons to distinct class 3 Semaphorins. *Neuron* **45**, 513–523 (2005).
- Gu, C. *et al.* Neuropilin-1 conveys semaphorin and VEGF signaling during neural and cardiovascular development. *Dev. Cell* **5**, 45–57 (2003).
- Giger, R. J. *et al.* Neuropilin-2 is required *in vivo* for selective axon guidance responses to secreted semaphorins. *Neuron* **25**, 29–41 (2000).
- Famiglietti, E. V. Jr & Kolb, H. Structural basis for ON- and OFF-center responses in retinal ganglion cells. *Science* **194**, 193–195 (1976).
- Viney, T. J. *et al.* Local retinal circuits of melanopsin-containing ganglion cells identified by transsynaptic viral tracing. *Curr. Biol.* **17**, 981–988 (2007).
- Zhang, D. Q. *et al.* Intraretinal signaling by ganglion cell photoreceptors to dopaminergic amacrine neurons. *Proc. Natl Acad. Sci. USA* **105**, 14181–14186 (2008).
- Hattar, S., Liao, H. W., Takao, M., Berson, D. M. & Yau, K. W. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science* **295**, 1065–1070 (2002).
- Pires, S. S. *et al.* Differential expression of two distinct functional isoforms of melanopsin (Opn4) in the mammalian retina. *J. Neurosci.* **29**, 12332–12342 (2009).
- Suto, F. *et al.* Interactions between plexin-A2, plexin-A4, and semaphorin 6A control lamina-restricted projection of hippocampal mossy fibers. *Neuron* **53**, 535–547 (2007).
- Stacy, R. C. & Wong, R. O. Developmental relationship between cholinergic amacrine cell processes and ganglion cell dendrites of the mouse retina. *J. Comp. Neurol.* **456**, 154–166 (2003).
- Kay, J. N. *et al.* Transient requirement for ganglion cells during assembly of retinal synaptic layers. *Development* **131**, 1331–1342 (2004).
- Runker, A. E., Little, G. E., Suto, F., Fujisawa, H. & Mitchell, K. J. Semaphorin-6A controls guidance of corticospinal tract axons at multiple choice points. *Neural Develop.* **3**, 34 (2008).
- Kerjan, G. *et al.* The transmembrane semaphorin *Sema6A* controls cerebellar granule cell migration. *Nature Neurosci.* **8**, 1516–1524 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K.-W. Yau for the C-terminal melanopsin antibody, F. Suto for the *PlexA4* antibody, B. Howell for the Dab-1 antibody, Y. Yoshida for the *PlexA1*^{-/-} eyes, P. Mombaerts for the *PlexB1*^{-/-} and *PlexB3*^{-/-} mice (unpublished), C. Gu for the *PlexD1*^{loxP}; *nestin cre* (unpublished) eyes and M. Tessier-Lavigne for the *PlexA4*^{-/-} mice. We also thank J. Nathans, S. Hattar, K. Mandai and M. Riccomagno for comments on the manuscript and discussions, and members of the Kolodkin laboratory for assistance. This work was supported by R01 NS35165 to A.L.K., a predoctoral fellowship from the Nakajima Foundation to R.L.M., the Fondation pour la Recherche Médicale (Programme équipe FRM) to A.C., the Fondation Retina France to K.T.N.-B.-C., and a PhD fellowship from the Paris School of Neuroscience (ENP) to A.P. A.L.K. is an investigator of the Howard Hughes Medical Institute.

Author Contributions R.L.M., A.C. and A.L.K. conceived and designed the experiments; R.L.M. performed most of the experiments and data analysis; K.T.N.-B.-C., A.P. and A.C. participated in the phenotypic analyses of *Sema6A* mutant mice and provided *PlexA2*^{-/-} and *PlexB2*^{-/-} mutants; T.C.B. performed cholera toxin injections and provided suggestions and reagents; R.L.M. and A.L.K. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.L.K. (kolodkin@jhmi.edu).

METHODS

Animals. The day of birth in this study is designated as postnatal (P) day 0. The *PlexA4*-deficient and *Sema6A* gene-trap mouse lines were previously described^{9,11}. The *PlexA1*^{-/-}, *PlexA2*^{-/-}, *PlexA3*^{-/-}, *PlexB2*^{-/-}, *PlexC1*^{-/-}, *Npn-1*^{Sema-/Sema-} and *Npn-2*^{-/-} mice were also described elsewhere^{11–13,19,24–27}. The *PlexB1*^{-/-} and *PlexB3*^{-/-} were obtained from P. Mombaerts (unpublished). The *PlexDI*^{-flox}; *nestin cre* mice were generated by G. Gu (unpublished).

Immunohistochemistry. Eyes were fixed in 4% paraformaldehyde for 1 h at 4 °C, equilibrated in 30% sucrose/PBS and embedded in OCT embedding media (Tissue-Tek). Retinal sections (20–40 µm) were blocked in 5% fetal bovine serum in 1× PBS and 0.4% Triton X-100 for 1 h at room temperature and then incubated overnight at 4 °C with primary antibodies: rabbit anti-tyrosine hydroxylase (Millipore at 1:1,000), sheep anti-tyrosine hydroxylase (Millipore at 1:400), rabbit anti-N-terminal melanopsin (ATS at 1:2,000), rabbit anti-C-terminal rat melanopsin (a gift from K.-W. Yau at 1:500)¹⁷, rabbit anti-calbindin (Swant at 1:2,500), rabbit anti-calretinin (Swant at 1:2,500), goat anti-calretinin (Swant at 1:2,500), rat anti-R-cadherin (Developmental Studies Hybridoma Bank at 1:200), goat anti-ChAT (Millipore at 1:100), rabbit anti-Dab-1 (a gift from B. Howell at 1:500), mouse anti-PKCα (Millipore at 1:200), mouse anti-Goz (Millipore at 1:500), goat anti-mouse *Sema6A* (R&D Systems at 1:200), Armenian hamster anti-*PlexA4* (a gift from F. Suto at 1:400)¹⁹, guinea pig anti-vGlut3 (Millipore at 1:2,500), chicken anti-β-gal (Abcam at 1:100), mouse anti-Brn3a (Millipore at 1:20), rabbit anti-Pax6 (Covance at 1:1,000) and goat anti-Chx10 (Santa Cruz at 1:50). Sections were washed six times for 5 min in 1× PBS and then incubated with secondary antibodies and TO-PRO-3 (Molecular Probes at 1:400) for 1 h at room temperature. Sections were washed six times for 5 min in PBS and coverslips were mounted using Vectashield HardSet Fluorescence Mounting Medium (Vector Laboratories), and confocal fluorescence images were taken using a Zeiss Axioskop2 Mot Plus, LSM 5 Pa confocal microscope. The regions we imaged did not include areas near the peripheral edges or the optic nerve head of retinas.

Wholemout retina staining. Enucleated eyes were fixed in 4% paraformaldehyde for 1 h at 4 °C. Whole retina cups were dissected out under a microscope and blocked in PBS containing 5% fetal bovine serum and 0.4% Triton X-100 for 2–3 h at room temperature. Retina cups were then incubated with primary antibodies in PBS containing 5% fetal bovine serum, 0.4% Triton X-100 and 20% dimethyl sulphoxide (DMSO) for 3–4 days at room temperature. Retinas were washed in PBS + 0.4% Triton X-100 for 7–8 h at room temperature and incubated with secondary antibodies in PBS containing 5% fetal bovine serum, 0.4% Triton X-100 and 20% DMSO for 24–36 h at room temperature. Retinas were washed in PBS + 0.4% Triton X-100 for 7–8 h at room temperature and flat mounted for confocal fluorescence images.

In situ hybridization. *In situ* hybridization was performed on either fresh frozen or PFA-fixed retina sections (20 µm thickness) as described previously¹³. Digoxigenin-labelled cRNA probes for *PlexA4* and *Sema6A* were used as previously described²⁸. Colourimetric *in situ* hybridization was in some cases followed by fluorescence immunohistochemistry and subsequent pseudocolouring of bright field images.

Quantification of anti-TH and anti-C-terminal melanopsin co-localized puncta. Confocal images of two selected regions (112 µm × 112 µm field) from

each retina (*n* = 3 retinas from three animals for wild-type and *PlexA4*^{-/-} genotypes) were double-immunostained with anti-TH and anti-C-terminal melanopsin, and the number of anti-TH-immunoreactive puncta co-localized with anti-C-terminal melanopsin in the S4/S5 sublaminae of the IPL was quantified.

Density recovery profile analysis. Density recovery profile analysis was performed as previously described^{6,29,30}. Confocal images of five selected regions (447 µm × 447 µm field) from each wholemount retina (*n* = 3 retinas from three animals for wild-type and *PlexA4*^{-/-} genotypes) were used to measure the density recovery profile of dopaminergic amacrine cells or ipRGCs. The retinal regions we used for this analysis did not include the areas near the peripheral edges or the optic nerve head.

X-gal staining. Eyes were fixed in 4% paraformaldehyde at 4 °C for 30 min, equilibrated in 30% sucrose/PBS, and embedded in OCT embedding media. Retina sections (20 µm) were stained with 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 2 mM MgCl₂ and 1 mg ml⁻¹ X-gal for 1–2 h at room temperature. Tissue sections were rinsed in PBS and bright-field images were taken.

Genetic interaction analysis. Retinal cross sections (40 µm thickness) from adult wild-type, *PlexA4*^{+/-}, *Sema6A*^{+/-} and *Sema6A*^{+/-}; *PlexA4*^{+/-} mice were immunostained with anti-TH (*n* = 4 retinas from four animals for each genotype). The number of anti-TH-immunoreactive puncta localized in the S4/S5 sublaminae of the IPL in five selected regions (149 µm × 149 µm field) from each retina was quantified. The retinal regions we used for this analysis did not include the areas near the peripheral edges or the optic nerve head.

Cholera toxin injection. Mice were anaesthetized with ketamine. Eyes were injected intravitreally with 1 µl of 2 mg ml⁻¹ cholera toxin B subunit solution conjugated with Alexa Fluor 488 or 594 (Invitrogen). Four to five days after injection, mice were perfused intracardially with 4% paraformaldehyde in PBS and brains were isolated. Brain sections (100 µm) were cut using a vibratome, and fluorescence images were taken.

Statistical analysis. Statistical differences for mean values among multiple groups were determined using Tukey's multiple comparison test. The criterion for statistical significance was set at *P* < 0.05. Error bars, s.e.m.

24. Yoshida, Y., Han, B., Mendelsohn, M. & Jessell, T. M. PlexinA1 signaling directs the segregation of proprioceptive sensory axons in the developing spinal cord. *Neuron* **52**, 775–788 (2006).
25. Cheng, H. J. et al. Plexin-A3 mediates semaphorin signaling and regulates the development of hippocampal axonal projections. *Neuron* **32**, 249–263 (2001).
26. Friedel, R. H. et al. Plexin-B2 controls the development of cerebellar granule cells. *J. Neurosci.* **27**, 3921–3932 (2007).
27. Pasterkamp, R. J., Peschon, J. J., Spriggs, M. K. & Kolodkin, A. L. Semaphorin 7A promotes axon outgrowth through integrins and MAPKs. *Nature* **424**, 398–405 (2003).
28. Suto, F. et al. Plexin-a4 mediates axon-repulsive activities of both secreted and transmembrane semaphorins and plays roles in nerve fiber guidance. *J. Neurosci.* **25**, 3628–3637 (2005).
29. Rodieck, R. W. The density recovery profile: a method for the analysis of points in the plane applicable to retinal studies. *Vis. Neurosci.* **6**, 95–111 (1991).
30. Rockhill, R. L., Euler, T. & Masland, R. H. Spatial order within but not between types of retinal neurons. *Proc. Natl Acad. Sci. USA* **97**, 2303–2307 (2000).

9p21 DNA variants associated with coronary artery disease impair interferon- γ signalling response

Olivier Harismendy^{1,2*}, Dimple Notani^{3*}, Xiaoyuan Song³, Nazli G. Rahim⁴, Bogdan Tanasa^{3,5}, Nathaniel Heintzman⁶, Bing Ren⁶, Xiang-Dong Fu⁷, Eric J. Topol⁴, Michael G. Rosenfeld³ & Kelly A. Frazer^{1,2,8}

Genome-wide association studies have identified single nucleotide polymorphisms (SNPs) in the 9p21 gene desert associated with coronary artery disease (CAD)^{1–4} and type 2 diabetes^{5–7}. Despite evidence for a role of the associated interval in neighbouring gene regulation^{8–10}, the biological underpinnings of these genetic associations with CAD or type 2 diabetes have not yet been explained. Here we identify 33 enhancers in 9p21; the interval is the second densest gene desert for predicted enhancers and six times denser than the whole genome ($P < 6.55 \times 10^{-33}$). The CAD risk alleles of SNPs rs10811656 and rs10757278 are located in one of these enhancers and disrupt a binding site for STAT1. Lymphoblastoid cell lines homozygous for the CAD risk haplotype show no binding of STAT1, and in lymphoblastoid cell lines homozygous for the CAD non-risk haplotype, binding of STAT1 inhibits *CDKN2BAS* (also known as *CDKN2B-AS1*) expression, which is reversed by short interfering RNA knockdown of *STAT1*. Using a new, open-ended approach to detect long-distance interactions, we find that in human vascular endothelial cells the enhancer interval containing the CAD locus physically interacts with the *CDKN2A/B* locus, the *MTAP* gene and an interval downstream of *IFNA21*. In human vascular endothelial cells, interferon- γ activation strongly affects the structure of the chromatin and the transcriptional regulation in the 9p21 locus, including STAT1-binding, long-range enhancer interactions and altered expression of neighbouring genes. Our findings establish a link between CAD genetic susceptibility and the response to inflammatory signalling in a vascular cell type and thus demonstrate the utility of genome-wide association study findings in directing studies to novel genomic loci and biological processes important for disease aetiology.

Genome-wide association studies (GWAS) have identified eight SNPs in the 9p21 interval strongly associated with CAD^{1–4} and other vascular diseases^{11,12}, all of which are highly correlated ($r^2 > 0.8$) and located in a 53-kb linkage disequilibrium block. The haplotype diversity in this interval is limited in Caucasians with ~25% of individuals homozygous for the risk haplotype that confers a ~2 fold greater risk of myocardial infarction than noncarriers². Independent studies have identified four more SNPs associated with type 2 diabetes (T2D) in an adjacent but distinct 11-kb linkage disequilibrium block^{5–7}. The associated CAD and T2D SNPs (Supplementary Table 1) lie in a gene desert (Fig. 1) flanked by *CDKN2B* (130 kb upstream) and *DMRTA1* (370 kb downstream), indicating that the functional variants underlying the association are probably in regulatory elements. The CAD interval overlaps with the 3' end of a non-coding gene with unknown function, *CDKN2BAS*, which is co-expressed with the *CDKN2A/B* locus¹³.

In this study we used a multi-pronged approach involving cellular assays and population sequencing to identify and functionally

characterize the variants underlying the 9p21 association with CAD (Supplementary Fig. 1). We initially sought to identify regulatory elements in the 9p21 gene desert by examining transcription factor binding and chromatin modification profiles in human cells including HeLa, K562 and human embryonic stem cells¹⁴. Histone H3 trimethylated at lysine 4 (H3K4me3) is associated with promoters of active genes, and by looking for this mark we determined that the *CDKN2B* and *DMRTA1* promoters were the only ones in the interval (Fig. 1). CTCF-binding sites mark insulators¹⁵; by analysing data of this factor in HeLa cells, we identified seven potential insulators in the 9p21 interval. One insulator is located close to the *CDKN2B* promoter and another one is located 130 kb upstream, in the CAD interval. Lastly, we searched for marks indicative of enhancers; enrichment of histone H3 monomethylation at lysine 4 (H3K4me1), binding of p300 and MED1, presence of DNase hypersensitivity sites (DHS) and depletion of H3K4me3¹⁶. Looking at these marks, we predict the locations of 33 enhancers, of which 26 are significantly enriched in conserved sequences ($P < 0.01$; Supplementary Table 2). Six enhancers are proximal to the CAD interval; nine enhancers are located in the CAD interval (referred to as ECAD1–9), two in the T2D interval (referred to as ET2D1–2) and 16 distal to the T2D interval. The majority of the 33 predicted enhancers fall in the proximal part of the gene desert, in or near the CAD and T2D intervals. These findings were further supported by the analysis of publicly available genome-wide data sets generated to predict regulatory elements using a variety of cell types (Fig. 1). Additionally, we validated the enhancer activity of the ECAD2, ECAD4, ECAD5, ECAD7, ECAD8, ECAD9 and ET2D1 elements using luciferase reporter assays in HeLa cells (data not shown). Interestingly, we determined that the 9p21 interval is the second densest gene desert for predicted enhancers (7.5 per 100 kb) in the human genome and the one containing the most disease-associated variants (10 SNPs; Supplementary Table 3). These data indicate that the 9p21 gene desert has an important regulatory function.

To identify the complete set of DNA variants in the 9p21 gene desert we sequenced a 196-kb interval (from the *CDKN2B* promoter to 65 kb distal to the T2D interval) in 50 individuals of European descent and used the variants to refine the pattern of linkage disequilibrium in the interval (Fig. 2a). We identified 765 variants (31 insertions and/or deletions (indels) and 734 SNPs) and used the r^2 correlation coefficient to identify variants in linkage disequilibrium with any of the 8 CAD- or 4 T2D-associated SNPs. Of the 765 variants, 41 spanning 44 kb are in perfect linkage disequilibrium with a CAD-associated SNP, increasing to 131 variants spanning 111 kb when lowering the r^2 threshold to 0.5 (Fig. 2b). In contrast to the CAD interval, only 23 variants were in linkage disequilibrium ($r^2 \geq 0.5$) with at least one of the 4 T2D-associated SNPs, spanning 11 kb. No variants were in linkage disequilibrium ($r^2 \geq 0.5$) with both CAD- and T2D-associated SNPs.

¹Department of Pediatrics and Rady's Children's Hospital, University of California at San Diego, School of Medicine, La Jolla, California 92093, USA. ²Moore's UCSD Cancer Center, University of California at San Diego, La Jolla, California 92093, USA. ³Department of Medicine, Howard Hughes Medical Institute, University of California at San Diego, School of Medicine, La Jolla, California 92093, USA. ⁴Scripps Genomic Medicine, Scripps Translational Science Institute, The Scripps Research Institute, 3344 N. Torrey Pines Court, La Jolla, California 92037, USA. ⁵Kellogg School of Science and Technology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. ⁶Ludwig Institute for Cancer Research, University of California at San Diego, School of Medicine, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁷Department of Cellular and Molecular Medicine, University of California at San Diego, School of Medicine, La Jolla, California 92093, USA. ⁸Institute for Genomic Medicine, University of California at San Diego, La Jolla, California 92093, USA.

*These authors contributed equally to this work.

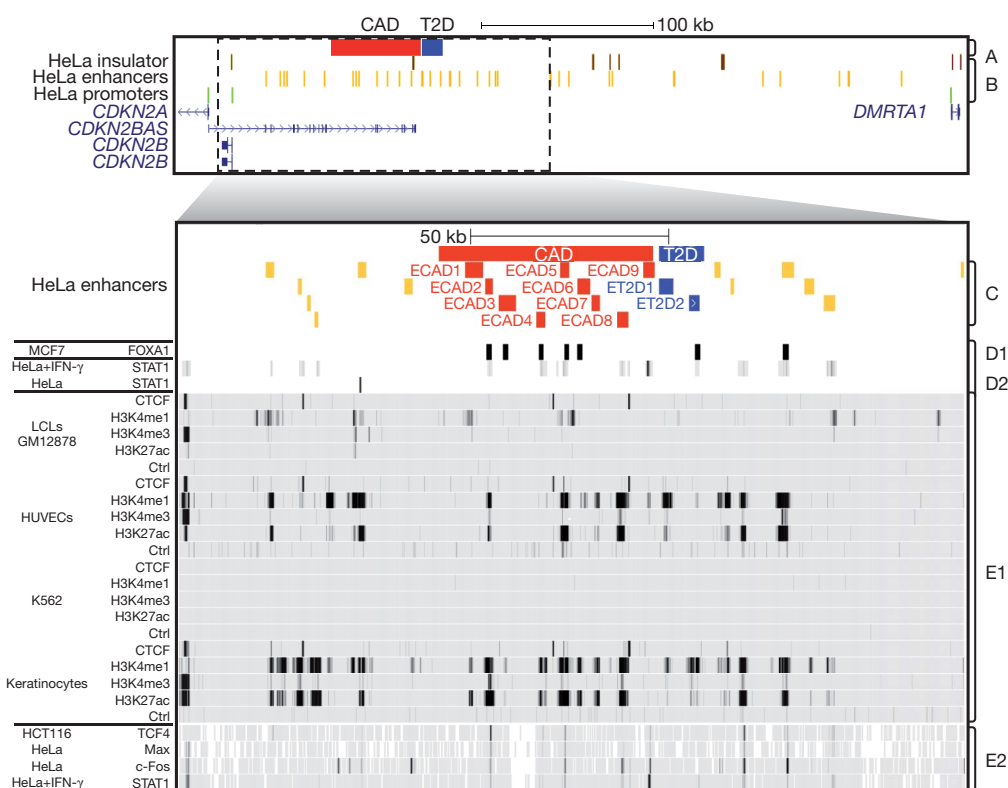


Figure 1 | Functional annotation of the 9p21 interval. The locations of the core CAD- and T2D-associated intervals (track A) and the predicted insulators (brown), enhancers (yellow) and promoters (green) in HeLa cells¹⁴ are indicated (track B). The enhancers are distributed (track C) between the CAD interval (red), the T2D interval (blue) or located outside (orange). The location

of the binding sites for FOXA1 in MCF7 cells (track D2)²⁹ and STAT1 in IFN- γ -treated and untreated HeLa cells (track D3)¹⁸ as well as the distribution of 9p21 chromatin marks in the ENCODE data³⁰ (tracks E1 and E2; Supplementary Methods) are indicated. Ctrl, control.

The functional variants underlying the association of 9p21 with CAD and T2D are highly likely to be a subset of the 131 and 23 candidate variants in linkage disequilibrium ($r^2 \geq 0.5$) with the initial CAD- or T2D-associated SNPs, respectively. However, the strong linkage disequilibrium structure in the 9p21 gene desert presents a hurdle for the precise identification of the causative functional variants using only genetic evidence. Thus, we determined which of the candidate variants are located in functional elements. Five of the CAD candidate variants are located in exons and 118 in introns of *CDKN2BAS*; however, none of the exonic variants are located in a conserved element, indicating that they are unlikely to be functional. Thirty-three CAD candidate variants are located in ECAD1–9 sequences; ECAD9 is significantly enriched for variants (11 in 1,700 bp; $P < 3.1 \times 10^{-5}$). The remaining 22 variants are distributed in ECAD1 (3 variants), ECAD2 (2), ECAD3 (8), ECAD4 (1), ECAD5 (3), ECAD6 (1), ECAD7 (3) and ECAD8 (1). Six of the 23 T2D candidate variants are located in a T2D enhancer, 5 of which are in ET2D1 and 1 in ET2D2.

To identify candidate regulatory variants, we computationally determined which variants disrupt consensus transcription-factor-binding sites in the predicted enhancers (Supplementary Tables 4 and 5). Although many of the enhancers in the CAD interval had variants disrupting transcription-factor-binding sites, ECAD9—containing eight such variants—was the most notable. Additionally, the SNP (rs10757278) most consistently associated with increased risk of CAD¹⁷ is in ECAD9 and disrupts a transcription-factor-binding site involved in the inflammatory response (STAT1; Supplementary Table 4). The rs10811656 and rs10757278 SNPs are located 4-bp apart in the predicted ECAD9 STAT1-binding site in the non-risk haplotype, which is disrupted in the risk haplotype (TTCTGGTAA > TTCTGGTAG). STAT1 has previously been shown to bind this locus in HeLa cells¹⁸, which are heterozygous for rs10811656/rs10757278 alleles, when treated with interferon- γ (IFN- γ). We observed the

binding of STAT1 at the same locus in human vascular endothelial cells (HUVECs) heterozygous for the rs10811656/rs10757278 alleles on treatment with IFN- γ (Fig. 3a). We examined the effect of IFN- γ treatment in heterozygous HeLa and HUVEC cells on the expression of *CDKN2B* and *CDKN2BAS*. Interestingly, we noticed that IFN- γ treatment leads to the repression and induction of *CDKN2B* and *CDKN2BAS*, respectively. This effect is greater in HUVECs, where *CDKN2BAS* expression is induced fourfold and *CDKN2B* transcription is repressed twofold (Fig. 3b). These results are consistent with the epigenetic transcriptional repression of *CDKN2B* induced by ANRIL, the transcript encoded by *CDKN2BAS*¹⁹.

To validate the disruption of STAT1 occupancy at the rs10811656/rs10757278 risk alleles, we used lymphoblastoid cell lines (LCLs) homozygous for the CAD risk haplotype. STAT1 is constitutively expressed at high levels in LCLs²⁰. We showed by ChIP that STAT1 binds at ECAD9 in two LCLs homozygous for the CAD non-risk haplotype (2.7-fold enrichment), whereas it does not bind in two LCLs homozygous for the risk haplotype (Fig. 3c). To establish further a functional link between STAT1 occupancy in ECAD9 and expression of *CDKN2BAS*, we assessed the expression of *CDKN2BAS* in LCLs on STAT1 short interfering RNA (siRNA)-mediated knockdown; the expression of *CDKN2BAS* was significantly up-regulated (sevenfold) in LCLs homozygous for the CAD non-risk haplotype, but the effect was quite small in LCLs homozygous for the CAD risk haplotype (Fig. 3d). These results are consistent with the fact that STAT1 fails to bind ECAD9 in the CAD risk haplotype and thus is unlikely to participate in *CDKN2BAS* regulation. Interestingly, these results indicate that the effects of STAT1 binding at ECAD9 may have cell-type-specific differences on the expression of *CDKN2BAS* between LCLs (repression) and HUVECs (activation) (Supplementary Discussion).

To determine which genes are interacting with and potentially regulated by the ECAD9 enhancers we examined all genes (20 upstream

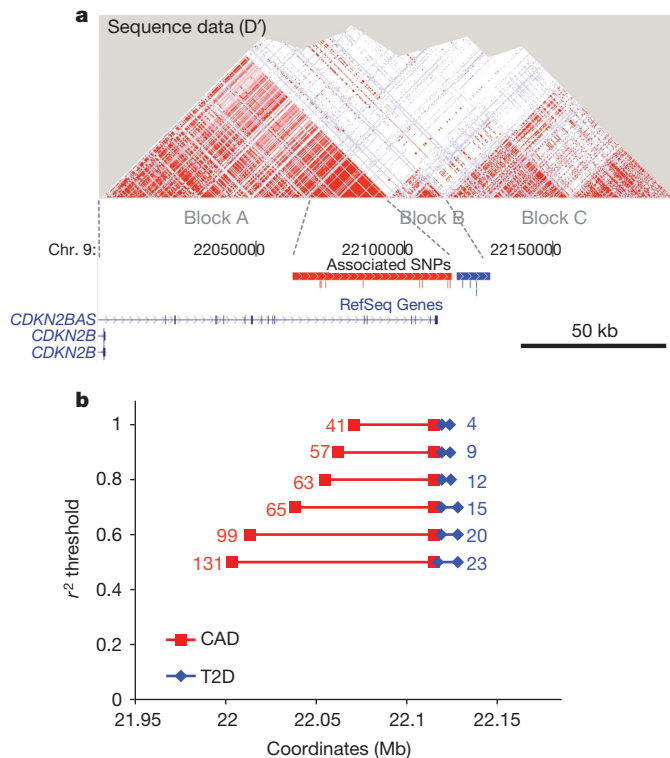


Figure 2 | Linkage disequilibrium analysis of the 9p21 interval. **a**, The sequenced interval shows three linkage disequilibrium blocks. The 53-kb CAD interval is located at the 3' end of the first block (A), the second block (B) spans 11 kb corresponding to the T2D interval and the third block (C) spans 63 kb. Linkage disequilibrium map (D') based on variants identified in the 50 sequenced samples. **b**, Number of variants in linkage disequilibrium at various r^2 thresholds (y axis) with any CAD- (red) or T2D-associated variants (blue). The distance spanned by the SNPs in linkage disequilibrium is indicated (x axis).

and 1 downstream) located within 2 Mb for interactions. As long-range enhancer interactions are often tissue specific²¹ we chose to use HUVEC cells as a model system for vascular endothelium. We performed chromatin conformation capture (3C) using cross-linked HUVECs grown in standard conditions, followed by BamH1 digestion of the chromatin and diluted re-ligation²². We analysed the 3C using a novel approach derived from combining 3C with DNA selection and ligation²³ (method referred to as 3D-DSL) and examined the sequences flanking the ligated acceptor and donor BamH1 sites using high-throughput sequencing. The resolution of 3D-DSL is limited by the

distribution of BamH1 sites, and to compensate for this we selected 6 BamH1 sites flanking ECAD9 and other neighbouring enhancers as acceptor sites, and 145 BamH1 donor sites distributed throughout the 2 Mb of sequence tested for interactions. We identified nine donor sites interacting strongly with the enhancer acceptor sites (Fig. 4a), of which seven are sufficiently distant (>45 kb) from the acceptor sites to be recognized as specific (Supplementary Table 6). They occur in four areas: (1) the *CDKN2A/B* locus; (2) the *MTAP* gene; (3) downstream of *IFNA21*; and (4) downstream of the T2D-associated region (DOTAR), an interval of unknown function. Accounting for the resolution limitations due to the distribution of BamH1 sites, we did not assign ligated acceptor sites as belonging to ECAD9 but rather refer to them as 'associated enhancers' and grouped the donor site interactions in *CDKN2A* and *CDKN2B* as 'CDKN2A/B'. We confirmed the interactions between the associated enhancers and surrounding genes identified by the 3D-DSL approach using traditional methods. The interaction between the associated enhancers and *IFNA21*, which spans 946 kb, was assayed by fluorescence *in situ* hybridization (FISH). We examined whether or not treatment with IFN- γ would affect this long-range interaction and observed that the enhancers and *IFNA21* are in close proximity in 58% versus 37% of the chromosomes analysed in HUVEC nuclei in the presence and absence of IFN- γ , respectively ($P < 1.63 \times 10^{-25}$; Supplementary Fig. 2 and Supplementary Table 7). These data indicate that this interaction occurs basally, but is significantly remodelled on treatment with IFN- γ . We validated more closely interacting loci via site-specific polymerase chain reaction (PCR) on the religated 3C DNA and verified that the interaction between the associated enhancers and the *CDKN2A/B* locus is strengthened in the presence of IFN- γ whereas the interaction with *MTAP* is reduced (Fig. 4b).

Our study shows that IFN- γ stimulation and STAT1 binding at ECAD9 have important roles in regulating the expression of *CDKN2B* and *CDKN2BAS*. Our findings are consistent with previous studies observing a correlation between CAD risk variants and *CDKN2A/B* expression in lymphocytes¹⁰, and reduced expression of *CDKN2A/B* in cardiac and other tissues in a mouse model with the orthologous CAD interval deleted⁸. We demonstrate also that the associated enhancers physically interact with the intervals encoding *CDKN2A/B*, *MTAP* and *IFNA21* in HUVECs and that these interactions are remodelled on IFN- γ treatment, thus indicating that the STAT1-binding site in ECAD9 is a key regulatory sequence. STAT1 is a downstream effector of the pathway that mediates response to inflammation, which is associated with angiogenesis²⁴ and atherosclerosis pathogenesis²⁵ in endothelial tissue. Thus, in endothelial tissues, the biological effects of the rs10811656 and rs10757278 risk alleles disrupting the ECAD9 STAT1-binding site might be augmented during activation of the inflammatory

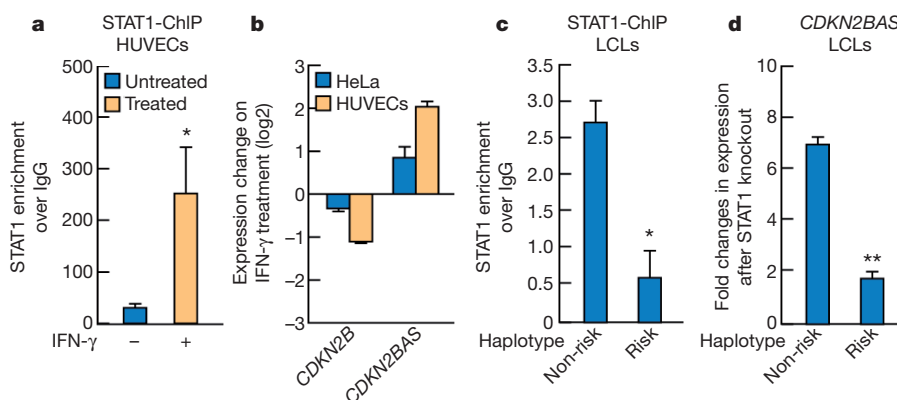


Figure 3 | In vivo effects of the ECAD9 variants. **a**, Enrichment of the ECAD9 STAT1-binding site by anti-STAT1 ChIP in HUVEC cells untreated or treated with IFN- γ . **b**, Changes in level of expression of *CDKN2B* and *CDKN2BAS* genes on treatment with IFN- γ in HeLa and HUVEC cells. **c**, Enrichment of the ECAD9 STAT1-binding site by anti-STAT1 ChIP in LCLs

homozygous for the CAD non-risk or CAD risk haplotypes. **d**, Expression level changes of *CDKN2BAS* in LCLs homozygous for non-risk or risk CAD haplotype after STAT1 knockdown by siRNA. Error bars represent the standard deviation over two or three replicate measurements (Supplementary Methods). * $P < 0.05$ and ** $P < 0.01$, using two-tailed Student's *t*-test.

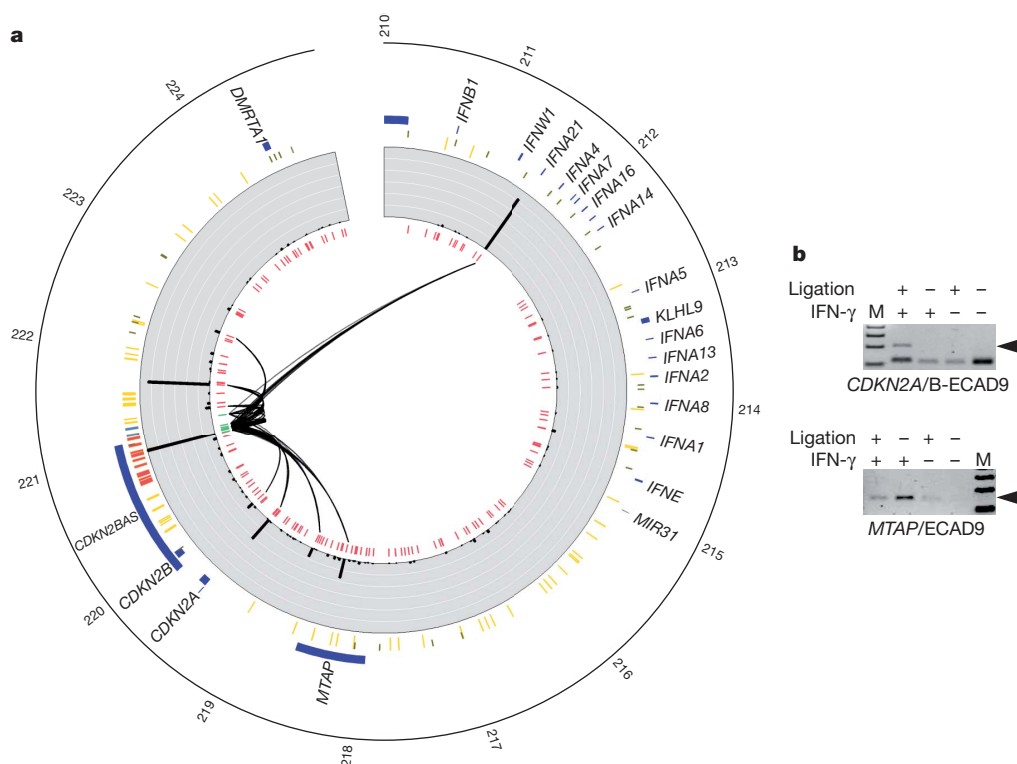


Figure 4 | Long-range interaction with the enhancer locus. **a**, Circular representation of the 3D-DSL results. The scale shows chromosome 9 position (hg18) in 100-kb increments. RefSeq genes (dark blue) and HeLa predicted enhancers (orange) are displayed with the CAD- (red) and T2D-associated (blue) enhancers. The histogram represents the normalized average number of

response. It is likely that the unique enhancer landscape of 9p21 is governed by higher-order chromatin structure and thus involved in different temporal- or tissue-specific expressions of additional genes than those identified in our study²⁶. Future studies assessing the effects of variants on the 9p21 chromatin landscape under different physiological conditions and cells types will undoubtedly provide further insights into the association of this interval with CAD and T2D susceptibility.

METHODS SUMMARY

HUVECs, LCLs and IFN- γ treatment. The two HUVEC cell lines used in the study were derived from male Caucasian donors genotyped as heterozygotes for the 9p21 CAD-associated SNPs. The four Epstein-Barr virus (EBV)-transformed LCLs were selected for their genotypes (two homozygous CAD risk or two homozygous CAD non-risk) using the HapMap data²⁷. Experiments were performed within 2–4 passages by treating cells with 100 ng ml⁻¹ of IFN- γ (R&D) for 24 h. Treated and untreated cells were subjected to PCR with reverse transcription (RT-PCR) (HUVECs), ChIP (HUVECs and LCLs), 3C analysis (HUVECs) or FISH (HUVECs).

Direct DSL and sequencing of 3C interactions. For probe design, we designed 12 acceptor probes in the interval chr9:22100523–22126469 (hg18; spanning from ECAD7 to the ET2D2 enhancer), on both strands immediately 3' of the six BamHI sites in the region. We designed 290 donor probes on both strands immediately 5' of the 145 BamHI sites in the interval chr9:21035934–22494089 (hg18; except where acceptor probes were designed). A universal sequence added to the probes is compatible with Illumina GA adapters for direct sequencing. The 12 acceptor probes and 290 donor probes (Supplementary Table 8) were pooled in equimolar amounts, separately.

For 3D-DSL sequencing, the DSL ligation products were prepared as described previously²³. 3C was performed as per ref. 28 and the products were sheared by sonication. The sheared DNA (200–600 bp) was purified and biotinylated. Donor and acceptor probes pools were annealed to the biotinylated 3C samples and the biotinylated DNA was bound on to streptavidin magnetic beads. The 5' phosphate of acceptor probes and 3' OH⁻ of donor probes were ligated using Taq DNA ligase. The ligated products were washed and eluted from the streptavidin

reads mapping to each BamHI donor site. The inner circle links connect BamHI acceptor sites for the nine most strongly interacting donor sites. **b**, PCR validation of the long-range interaction between ECAD9 and CDKN2A/B (top) or MTAP (bottom). Arrow indicates the specific product.

magnetic beads, followed by PCR amplification and deep sequencing on the Illumina GA2 (Supplementary Information).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 November 2009; accepted 16 December 2010.

- McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
- Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493 (2007).
- Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
- Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
- Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40**, 638–645 (2008).
- Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409–412 (2010).
- Jarinova, O. *et al.* Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler. Thromb. Biol.* **29**, 1671–1677 (2009).
- Liu, Y. *et al.* *INK4/ARF* transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. *PLoS ONE* **4**, e5027 (2009).
- Helgadottir, A. *et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genet.* **40**, 217–224 (2008).
- Newton-Cheh, C. *et al.* A common variant at 9p21 is associated with sudden and arrhythmic cardiac death. *Circulation* **120**, 2062–2068 (2009).
- Pasman, E. *et al.* Characterization of a germ-line deletion, including the entire *INK4/ARF* locus, in a melanoma-neural system tumor family: identification of *ANRIL*, an antisense noncoding RNA whose expression coclusters with *ARF*. *Cancer Res.* **67**, 3963–3969 (2007).
- Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

15. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
16. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
17. Tousoulis, D., Briasoulis, A., Papageorgiou, N., Antoniadis, C. & Stefanadis, C. Candidate gene polymorphisms and the 9p21 locus in acute coronary syndromes. *Trends Mol. Med.* **14**, 441–449 (2008).
18. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
19. Yap, K. L. *et al.* Molecular interplay of the noncoding RNA *ANRIL* and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of *INK4a*. *Mol. Cell* **38**, 662–674 (2010).
20. Najjar, I. *et al.* Latent membrane protein 1 regulates STAT1 through NF- κ B-dependent interferon secretion in Epstein-Barr virus-immortalized B cells. *J. Virol.* **79**, 4936–4943 (2005).
21. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with *MYC*. *Proc. Natl Acad. Sci. USA* **107**, 9742–9746 (2009).
22. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
23. Kwon, Y. S. *et al.* Sensitive ChIP-DSL technology reveals an extensive estrogen receptor α -binding program on human gene promoters. *Proc. Natl Acad. Sci. USA* **104**, 4852–4857 (2007).
24. Battle, T. E., Lynch, R. A. & Frank, D. A. Signal transducer and activator of transcription 1 activation in endothelial cells is a negative regulator of angiogenesis. *Cancer Res.* **66**, 3649–3657 (2006).
25. Ross, R. Atherosclerosis is an inflammatory disease. *Am. Heart J.* **138**, S419–S420 (1999).
26. Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403–413 (2006).
27. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
28. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
29. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
30. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank X. Wang, K. Trevarthen and M. Nakano for experimental assistance. We thank the Scripps Genomic Medicine Clinical team for sample collection. This work was partially supported by National Center for Research Resources grants 1U54RR025204, 1UL1RR025774 and 1UL1RR031980-01 and National Institutes of Health grants HL065445, DK39949, DK018477, DK074868, L65445, CA97134, DK74686, NS34934 and 1R21CA152613-01, the Department of Defence grant BC075019 and the Prostate Cancer Foundation. M.G.R. is an Investigator with the Howard Hughes Medical Institute.

Author Contributions O.H., D.N., B.R., E.J.T., K.A.F. and M.G.R. designed the experiments. D.N., X.S., N.G.R., N.H. and X.-D.F. carried out the experiments. O.H., D.N., X.S., B.T., K.A.F. and M.G.R. contributed to analysing the data. O.H., D.N., K.A.F. and M.G.R. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.J.T. (etopol@scripps.edu), M.G.R. (mrosenfeld@ucsd.edu) or K.A.F. (kafrazer@ucsd.edu).

METHODS

Annotation of the 9p21 gene desert. The two protein coding genes *CDKN2B* and *DMRTA1* flank the 9p21 gene desert. The CAD and T2D intervals (Fig. 2a) are located 130 kb upstream of *CDKN2B*. The CAD interval has a 48 kb overlap with the 3' end of a recently annotated non-coding gene with unknown function^{13,31}, *CDKN2BAS*, which is co-expressed with the *CDKN2A/B* locus¹³. Similar to many long non-coding RNA genes^{32,33}, *CDKN2BAS* lacks strong sequence conservation across its transcribed region but does contain some short strongly conserved elements that are probably the functional domains in the gene. Its transcriptional start site is ~300 bp from the *CDKN2A* promoter on the anti-sense strand, indicating a possible regulatory role³⁴.

Sequenced samples. Two-hundred and forty-four self-reported Caucasian males enrolled in the Scripps Translational Science Institute (<http://www.stsiweb.org/>) GeneHeart cohort were genotyped at ten SNPs (Supplementary Table 1) on the Sequenom MassARRAY platform according to the manufacturer's instructions. The eight CAD Sequenom genotypes (Supplementary Table 9) or four T2D genotypes derived from the sequence data were phased in haplotypes using the program PHASE 2.1.1³⁵. For deep sequencing of the 9p21 interval we selected 25 samples homozygous for the CAD risk haplotype (22 H1/H1 and 3 H1/H5), 24 samples homozygous for the CAD non-risk haplotype (15 H2/H2, 7 H2/H3, 1 H3/H3, 1 H2/H6), and 1 sample with a non-risk and a mixed haplotype (1 H2/H17) (Supplementary Table 10). Our selection of the 50 samples for sequencing considered some of the haplotype diversity in the population, without impairing our ability to assign newly identified variants to either the CAD risk or non-risk haplotypes. The sequencing method is described as Supplementary Information.

T2D- and CAD-associated variants. Using the Caucasian HapMap (version 24) genotypes, we identified 13 and 44 variants in linkage disequilibrium at an $r^2 \geq 0.5$ with any of the initial four T2D variants (rs7020996, rs2383208, rs10811661, rs10757283) or the initial eight CAD variants (Supplementary Tables 1, 4 and 5) respectively. Using the sequence data, we identified 20 and 131 variants in linkage disequilibrium at an $r^2 \geq 0.5$ with any of the initial three T2D variants (rs2383208, rs10811661, rs10757283) and eight initial CAD variants, respectively. One of the initial associated T2D SNPs (rs7020996) is located in a long-range PCR primer site that was part of the set of primers used to amplify the 196-kb interval and was thus filtered out in our initial analysis. Using the union of the two data sets (HapMap and sequencing), we identify a total of 23 and 131 variants in linkage disequilibrium at an $r^2 \geq 0.5$ with any of the four initial T2D variants and eight initial CAD variants, respectively, including the associated SNPs themselves (Fig. 2b).

ChIP-chip experiments. HeLa cells were obtained from ATCC. ChIP, DNA purification and ligation-mediated PCR were performed as described¹⁶ using commercially available (anti-H3K27ac, Abcam ab4729; anti-H3K4me1, Abcam ab8895; anti-H3K4me3, Upstate 07-473; anti-p300, Santa Cruz sc-585; anti-MED1, Santa Cruz sc-5334) or previously described antibodies^{15,16} and the ChIP DNA was hybridized to tiling microarrays and to custom-made enhancer microarrays (NimbleGen Systems) as described^{15,16}. ChIP-chip targets for CTCF, p300 and MED1 were selected with the Mpeak program³⁶. We used MA2C to normalize and call peaks on Nimblegen HD2 arrays³⁷. Enhancers were predicted as previously described³⁸. H3K4me3 is associated with promoters of active genes¹⁶, CTCF-binding sites mark insulators³⁹. Marks indicative of enhancers are enrichment of H3K4me1, binding of p300 and MED1, presence of DNase hypersensitivity sites and depletion of H3K4me3.

Enhancer density in gene deserts. We used the genomic coordinates (NCBI36) of the protein-coding genes⁴⁰ to identify gene-desert intervals longer than 400 kb in size. Using the coordinates of the predicted transcriptional enhancers in HeLa¹⁴ we counted the number of enhancers per interval and inferred the expected number of enhancers based on size and assuming homogeneous distribution. We tested for heterogeneity using a chi-squared test to compare the whole genome distribution and corrected for the number of tests using the Bonferroni correction. The gene deserts were then filtered for a significant enrichment in predicted enhancers (corrected P value < 0.01) and ranked by decreasing enhancer density (Supplementary Table 3).

For each of the selected gene deserts, we identified and counted the SNPs in association with diseases using the public GWAS catalogue (October 2009)⁴¹. Using the method described earlier in the genome assembly hg18, we identified 1,155 gene deserts of which 129 were significantly enriched in predicted enhancer elements (corrected P value < 0.01). Fifty of those harbour SNPs associated with diseases⁴¹. The 9p21 interval is the second densest gene desert for predicted enhancers (7.5 per 100 kb) and the one containing the most disease-associated variants (10 SNPs) (Supplementary Table 3). The density of enhancers in the 9p21 interval is thus six times higher than the genome-wide density (1.2 enhancers per 100 kb; $P < 6.54 \times 10^{-33}$).

Sequence conservation of enhancer elements. Conserved bases were defined as nucleotides with a conservation score ≥ 1.65 (95th percentile in the interval) in the multispecies sequence comparison track at UCSC (44 ways placental mammals

PhyloP conservation score). Each 9p21 predicted enhancer was tested to determine if it was significantly enriched for conserved sequences as follows: the ratio of the number of conserved bases to the total predicted enhancer length was compared to that of non-enhancer sequences in the re-sequenced interval using a chi-squared test. Enhancers with a P value < 0.01 were considered to be significantly enriched for conserved sequences (Supplementary Table 2). Of note, the predicted enhancers have already been shown to be statistically more evolutionary conserved¹⁴. However, owing to the ~1–2-kb resolution of the enhancer prediction method, we manually adjusted the boundaries of the enhancers located in CAD and T2D intervals (Supplementary Table 2) to include adjacent conserved sequences.

Identification of transcription-factor-binding sites. MotifLocator^{42,43} at a stringency of 0.01% was used to predict potential binding sites (transcription-factor-binding site models from the JASPAR and TRANSFAC databases) in the 60 bp of sequence upstream and downstream of each variant locus, both alleles were considered. We retained the transcription-factor-binding site predictions that were disrupted by one of the two alleles for all candidate variants in linkage disequilibrium ($r^2 \geq 0.5$) with CAD or T2D variants (Supplementary Tables 4 and 5).

Cell culture of HUVECs and LCLs and IFN- γ treatment. The two HUVEC cell lines (Lonza) used in the study were from male Caucasian donors and genotyped as heterozygotes for the 9p21 CAD-associated SNPs using the method described for HeLa genotyping (see Supplementary Methods). The EBV-transformed LCLs (Coriell Institute) were selected for their genotypes using the HapMap data²⁸: LCLs NA12156 and NA10839 were homozygous for the CAD risk haplotype, whereas LCLs NA12750 and NA10847 were homozygous for the non-risk haplotype. Experiments were performed within 2–4 passages by treating cells with 100 ng ml^{-1} of IFN- γ (R&D) for 24 h then subjected to RT-PCR (HUVECs), ChIP (HUVECs and LCLs), 3C analysis (HUVECs) or FISH (HUVECs), as described later.

ChIP. ChIP was performed essentially as described previously⁴⁴. In brief, cells were crosslinked for 10 min and were then subjected to sonication using Bioruptor (Diagenode) to fragment the chromatin to obtain 200–1,000-bp fragments. Sonicated chromatin was precleared with a cocktail containing 50% protein A/G beads slurry (Pierce), salmon sperm DNA and bovine serum albumin (BSA). Precleared chromatin was incubated with anti-STAT1 antibody (IP sample; SantaCruz biotechnology) or with rabbit-IgG (Upstate Biotechnology; IP control). Protein A/G bead cocktail was then added to pull-down the antibody-bound chromatin and was subjected to elution using sodium bicarbonate buffer containing SDS and dithiothreitol (DTT). Eluted chromatin was de-crosslinked and proteins were removed by treating with proteinase K. The purified immunoprecipitated chromatin was subjected to PCR amplification of specific targets using oligonucleotides primers (Supplementary Table 11) along with input chromatin and mock IP (IgG) control. The level of enrichment was measured using quantitative PCR with Q-Master mix (Agilent) on real-time thermocycler MX3000P (Startagene). The quantity of target specific IP DNA was first normalized to the input ($2^{-(Ct_{IP} - Ct_{input})}$). The level of enrichment was then calculated by dividing by the normalized quantity from the mock IP control (IgG). The average of three replicates (Fig. 3a) or three measurements in two cell lines (Fig. 3c) is shown. The significance was assessed using a two-tailed Student's t -test.

RT-PCR. RNA was isolated from IFN- γ -treated and untreated HUVEC and LCL cells and 1 μg total RNA was reverse transcribed using SuperScript III Reverse Transcriptase (Invitrogen) as per manufacturer's instructions. Quantitative PCRs were performed in MX3000P (Stratagene) using Q-PCR master mix (Agilent). For normalization, ΔCt values were calculated using the formula: $\Delta Ct = (Ct_{\text{target}} - Ct_{\text{control}})$, where control corresponds to the level of GAPDH transcript. Fold differences in normalized gene expression were calculated by dividing the level of expression of the treated sample with the untreated sample (Fig. 3b) or between siSTAT1 and siControl transfected cells (Fig. 3d). The result of three replicates (Fig. 3b) or three replicates in two cell lines (Fig. 3d) was averaged and the significance of changes was assessed using a two-tailed Student's t -test (Fig. 3d only).

siRNA analysis. Five million LCL cells either homozygous for the non-risk (NA12750 and NA10847) or for the risk (NA12156 and NA10839) CAD haplotype were electroporated with siSTAT1 (Ambion) and control siRNA (Ambion) using Amaxa Nucleofector I electroporator (AMAXA Biosystems). Cells were subjected to RT-PCR and ChIP analysis 24 h after electroporation as described earlier. The knockdown efficiency was verified by ChIP (Supplementary Fig. 3).

3C. 3C was performed as described previously²⁹. Briefly, 25 million cells were fixed by adding 1% formaldehyde at room temperature (22 °C) for 10 min. The reaction was stopped by adding glycine. Lysis buffer (500 μl of 10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% Igepal CA630; protease inhibitors (Sigma)) was added and cells were incubated on ice. Next, cells were lysed with a Dounce homogenizer by moving the pestle up and down ten times, incubating on ice for one minute followed by ten more strokes with the pestle. The suspension was spun down at 5,000 r.p.m. at room temperature. The supernatant was discarded and the pellet was washed twice with 500 μl ice-cold $1 \times$ NEBuffer 3 (NEB). The pellet was then

resuspended in 1× NEBuffer 3 and split into five 50-μl aliquots. Chromatin was subsequently digested overnight by adding 400 Units BamHI (NEB). Each digested chromatin mixture was ligated by adding T4 DNA Ligase (800 U) in 20 times the initial volume for 4 h at 16 °C. One aliquot out of five was kept as an unligated control where ligase was omitted. After incubation at 16 °C, the chromatin was de-crosslinked overnight at 65 °C and purified twice with phenol and then with phenol:chloroform (25:24). DNA was precipitated and pellets were air-dried before resuspending in 25 μl 1× TE (10 mM Tris, 1 mM EDTA, pH 8). To degrade any carryover RNA, 1 μl RNase A (1 mg ml⁻¹) was added to each tube and incubated at 37 °C for 15 min. Chromatin was used for subsequent DSL analysis (see below) or for validation PCR (primers are listed in Supplementary Table 11).

DSL and sequencing of 3C interactions. For probe design, we designed 12 acceptor probes in the interval chr9:22100523–22126469 (hg18; spanning from ECAD7 to the ET2D2 enhancer), on both strands immediately 3′ of the six BamHI sites in the region. The acceptor probes are 5′-pTCC-(locus)-CCTGTGGT CGTAGCATCAGC-3′, where (locus) indicates a 17-nucleotide sequence immediately 3′ of the BamHI site and CCTGTGGTCGTAGCATCAGC is the complementary sequence to the adaptor (Primer B-AD in Supplementary Table 11). We designed 290 donor probes on both strands immediately 5′ of the 145 BamHI site in the interval chr9:21035934–22494089 (hg18; except where acceptor probes were designed). The probe sequences are shown in Supplementary Table 8. The donor probes are 5′-AATGATACGGCGACCCAGAT-(locus)-GGA-3′, where (locus) indicates the 17-nucleotide sequences immediately 5′ of the BamHI site. The uniqueness of the (locus) oligonucleotides was verified with BLAST and BOWTIE alignment. The universal sequences above are adapted for the Illumina GA adapters and compatible with flow-cell bridge amplification. The 12 acceptor probes and 290 donor probes were pooled in equimolar amounts, separately.

For 3D-DSL sequencing, the DSL ligation products were prepared as described previously²³. The 3C products from HUVEC cells were sheared by sonication. The sheared DNA (200–600 bp) was purified from gel and subjected to biotinylation before precipitation with 100% ethanol. Donor and acceptor probe pools (final 2.5 fmol each probe) were annealed to the biotinylated 3C samples at 45 °C for 2 h followed by 10 min at 95 °C. Unbound oligonucleotides were removed by washes while biotinylated DNA was bound on to streptavidin magnetic beads. The 5′-phosphate of acceptor probes and 3′-OH of donor probes were ligated using Taq DNA ligase at 45 °C for 1 h. These ligated products were washed and eluted from beads and then amplified by PCR using primers A and B-AD (or Primer B-BC1 and -BC2 if bar coding was used; see Supplementary Table 11) for deep sequencing on the Illumina GA2 using Primer A as sequencing primer.

For 3D-DSL analysis, we first built a virtual library of the 3,480 possible donor-acceptor sequences by *in silico* concatenating all 12 acceptor sequences with 290 donor sequences. Because the religation products span a BamHI restriction site, we first selected sequence reads containing intervening GGATCC sequence (BamHI), aligned these reads to the virtual library of DSL donor-acceptor sequences using BOWTIE⁴⁵ allowing for no mismatches, and then counted the number of reads mapped to every interaction. The summary results of these steps are presented in Supplementary Table 12. The fraction of total reads mapped per donor site is presented in Supplementary Table 6. Although the most highly covered donor sites are consistent between the technical replicate samples, we observed some variability. Except for some low covered potential false-positive interactions, we think that the variability could be due to the stochastic nature of the long-range chromatin interactions and of the 3C assay itself. To strengthen our confidence in the results we averaged the two experiments and focused our study on donor sites covered by more than 1% of the reads on average over the two experiments (highlighted in bold in Supplementary Table 6). We realize that this is an empirical threshold and that some of the interactions below this threshold will also probably be real. Because of the low resolution and the stochastic nature of the 3D-DSL, constrained by the distribution of BamHI sites, we did not distinguish between sites located within 50 kb of each other (for example, grouping interactions in *CDKN2A* and *CDKN2B* as '*CDKN2A/B*' or ECAD7-9 and ET2D1-2 as 'associated enhancers').

For 3D-DSL validation by PCR, to validate the interactions between the enhancer interval and the *CDKN2A/B* and *MTAP* loci (Fig. 4b), we performed PCR on the 3C ligation product. We designed two sets of primer pairs (Supplementary Table 11): (1) between a BamHI donor site close to *CDKN2A/B* (hg18, chr9:21985877–21985882) and a BamHI acceptor site located in ECAD8 (hg18; chr9:22108031–22108036); (2) between a BamHI donor site in the *MTAP* gene (hg18; chr9:21831910–21831915) and a BamHI acceptor site located in ET2D2 (hg18; chr9:22126466–22126471). These experiments were all performed in replicate with the same effects observed. The identified long-range interaction between the

CDKN2A/B locus and the associated enhancers in untreated HUVECs was observed consistently by 3D-DSL, but not clearly identified by traditional, PCR-based 3C. This indicates that the 3D-DSL is a more sensitive method for the detection of long-range interactions. The interacting donor sites located in very close proximity to the associated enhancers cannot be resolved using 3C and were not assayed nor were the interactions with DOTAR because of the inability to design primers due to low GC content.

FISH. HUVEC cells were grown on polylysine-coated cover slips and treated with or without IFN-γ (100 ng ml⁻¹) for 24 h. Cells were fixed with freshly made 4% paraformaldehyde/PBS at room temperature and permeabilized with 0.5% Triton X-100. Freeze/thaw cells in liquid N₂, 20 s × 5 times and incubated in 0.1 N HCl for 5 min. To further permeabilize the cells, they were treated with 0.01 N HCl/0.002% pepsin for 3 min 40 s, stopped by 50 mM MgCl₂/PBS and equilibrated in 50% formamide/2× SSC for 2 h before hybridization. Two microlitres each of the two labelled BAC probes (Empire Genomics) were mixed with 7 μl of hybridization buffer (Empire genomics). The slides were placed face down on the hybridization mix, sealed with fixogum and air-dried. After denaturation at 76 °C for 3 min, the slides were transferred to a prewarmed dark wet box at 37 °C and hybridized overnight. The slides were then washed 3 times with 50% formamide/2× SSC/0.1% Tween 20 at 37 °C; 3 times with 0.1× SSC/0.1% Tween 20 at 60 °C; 2 times with 2× SSC, once at 60 °C and once at room temperature; rinsed with 1× PBS twice at room temperature. The cells were then mounted on the slides with prolong gold-DAPI antifade mounting reagent (Invitrogen), and analysed under fluorescent microscope counting at least 100 cells for each experiment. The experiments were done on three biological replicates. The probes used spanned the enhancer region (hg18; chr9:22100296–22249612; RP11-248B1 labelled with fluorescein; Empire Genomics) and the IFNA21 region (chr9:20996400–21158464; 942 kb away from the enhancer; RP11-113D19, labelled with 5-ROX; Empire Genomics). For each experiment, we counted the number of cells with biallelic, monoallelic or negative interactions. We then performed a chi-squared test for goodness of fit using these three values comparing between IFN-γ treated and untreated. We found that 58% of the chromosomes in 315 HUVEC nuclei analysed (heterozygous for the CAD risk and non-risk haplotypes) showed an overlapping signal from the two probes (Supplementary Fig. 2) when the cells were treated with IFN-γ, in contrast with 37% of overlapping chromosomes in the absence of IFN-γ treatment. These data indicate that this interaction occurs basally, but is significantly remodelled on treatment with IFN-γ ($P < 1.63 \times 10^{-25}$) (Supplementary Table 7).

- Broadbent, H. M. *et al.* Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the *ANRIL* locus on chromosome 9p. *Hum. Mol. Genet.* **17**, 806–814 (2008).
- Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5 (2006).
- Nesterova, T. B. *et al.* Characterization of the genomic *Xist* locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11**, 833–849 (2001).
- Yu, W. *et al.* Epigenetic silencing of tumour suppressor gene *p15* by its antisense RNA. *Nature* **451**, 202–206 (2008).
- Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
- Zheng, M., Barrera, L. O., Ren, B. & Wu, Y. N. ChIP-chip: data, model, and analysis. *Biometrics* **63**, 787–796 (2007).
- Song, J. S. *et al.* Model-based analysis of two-color arrays (MA2C). *Genome Biol.* **8**, R178 (2007).
- Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
- Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387–396 (1999).
- Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
- Hindorf, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A *Catalog of Published Genome-Wide Association Studies* (National Human Genome Research Institute, accessed on 30 October 2009).
- Thijs, G. *et al.* A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* **9**, 447–464 (2002).
- Gilchrist, M. *et al.* Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* **441**, 173–178 (2006).
- Solomon, M. J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for *in vivo* chromatin structures. *Proc. Natl Acad. Sci. USA* **82**, 6470–6474 (1985).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression

Zhihu Ding^{1,2,3,4}, Chang-Jiun Wu^{1,2,3,4*}, Gerald C. Chu^{1,2,5*}, Yonghong Xiao^{1,2}, Dennis Ho^{1,2,3,4}, Jingfang Zhang⁶, Samuel R. Perry^{1,2}, Emma S. Labrot^{1,2}, Xiaoqiu Wu^{2,7}, Rosina Lis^{2,7}, Yujin Hoshida^{8,9}, David Hiller¹⁰, Baoli Hu^{1,2}, Shan Jiang^{1,2}, Hongwu Zheng^{1,2,3,4}, Alexander H. Stegh^{1,2,3,4}, Kenneth L. Scott^{1,2,3,4}, Sabina Signoretti¹¹, Nabeel Bardeesy¹², Y. Alan Wang^{1,2}, David E. Hill^{3,13}, Todd R. Golub^{8,9}, Meir J. Stampfer^{15,16,17}, Wing H. Wong¹⁰, Massimo Loda^{2,5,7}, Lorelei Mucci^{15,17}, Lynda Chin^{1,2,3,4,14} & Ronald A. DePinho^{1,2,3,4}

Effective clinical management of prostate cancer (PCA) has been challenged by significant intratumoural heterogeneity on the genomic and pathological levels and limited understanding of the genetic elements governing disease progression¹. Here, we exploited the experimental merits of the mouse to test the hypothesis that pathways constraining progression might be activated in indolent *Pten*-null mouse prostate tumours and that inactivation of such progression barriers in mice would engender a metastasis-prone condition. Comparative transcriptomic and canonical pathway analyses, followed by biochemical confirmation, of normal prostate epithelium versus poorly progressive *Pten*-null prostate cancers revealed robust activation of the TGFβ/BMP–SMAD4 signalling axis. The functional relevance of SMAD4 was further supported by emergence of invasive, metastatic and lethal prostate cancers with 100% penetrance upon genetic deletion of *Smad4* in the *Pten*-null mouse prostate. Pathological and molecular analysis as well as transcriptomic knowledge-based pathway profiling of emerging tumours identified cell proliferation and invasion as two cardinal tumour biological features in the metastatic *Smad4*/*Pten*-null PCA model. Follow-on pathological and functional assessment confirmed cyclin D1 and SPPI as key mediators of these biological processes, which together with PTEN and SMAD4, form a four-gene signature that is prognostic of prostate-specific antigen (PSA) biochemical recurrence and lethal metastasis in human PCA. This model-informed progression analysis, together with genetic, functional and translational studies, establishes SMAD4 as a key regulator of PCA progression in mice and humans.

Adenocarcinoma of the prostate (PCA) is the most common form of cancer and the second leading cause of cancer death in American men². Current methods of stratifying tumours to predict outcome are based on clinical-pathological factors including Gleason grade, PSA and tumour stage³. These parameters are widely considered inadequate, which has motivated the genetic and biological study of PCA progression with the goal of identifying progression risk biomarkers capable of improving patient management⁴.

Genetic studies of human PCA has identified signature pathogenetic events⁵, a number of which have been validated and mechanistically defined in genetically engineered mouse models of PCA⁶. Prostate-specific *Pten* deletion (*Pten*^{pc-/-}) results in prostate intraepithelial neoplasia (PIN) which, following a long latency, can progress to

high-grade adenocarcinoma, albeit with minimally invasive and metastatic features^{7–10}. To understand this feeble progression phenotype, we conducted transcriptome comparison of *Pten*^{pc-/-} PIN relative to wild-type prostate epithelium (Supplementary Data 1). In addition to the expected PI3K and p53 (also known as TRP53) pathway representation⁸, knowledge-based pathway analysis revealed prominent TGFβ/BMP signalling in *Pten*^{pc-/-} PIN (Supplementary Fig. 1). Immunohistochemical and western blotting analyses of Smad4 expression confirmed robust increase in *Pten*^{pc-/-} PIN compared to wild-type prostate epithelium (Fig. 1a, b). In line with reported down-regulated expression of SMAD4 in a subset of human primary prostate tumours¹¹, Oncomine expression analysis showed consistent *SMAD4* downregulation in human PCA metastasis (Fig. 1c and Supplementary

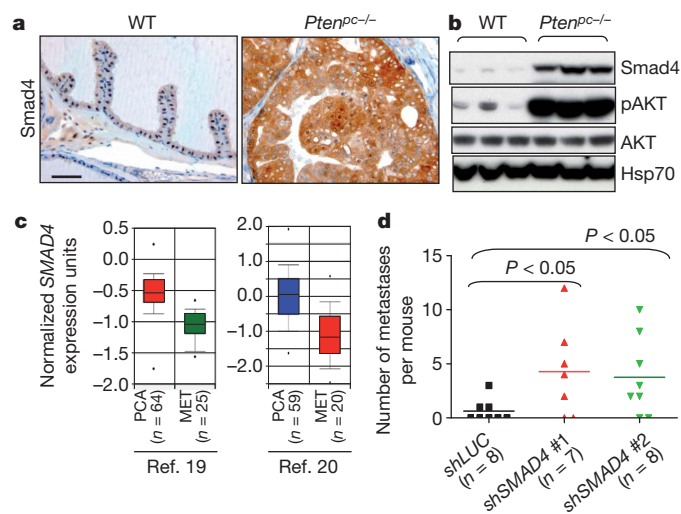


Figure 1 | SMAD4 is a putative suppressor of prostate tumour progression. **a, b**, Immunohistochemical (**a**) and western blot analysis (**b**) of wild-type (WT) and *Pten*^{pc-/-} mouse prostate tissues. Scale bar, 50 μm. **c**, Oncomine boxed plot of *SMAD4* expression levels between human PCA and metastasis in multiple data sets including those from ref. 19 and ref. 20. **d**, *SMAD4* knockdown enhanced metastatic potential to lung from PC3 cells implanted in renal capsule of immunocompromised nude mice.

¹Belfer Institute for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ²Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ³Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ⁶McArdle Lab for Cancer Research, University of Wisconsin, Madison, Wisconsin 53706-1526, USA. ⁷Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ⁸Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ⁹The Eli and Edythe L. Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. ¹⁰Department of Statistics, Stanford University, Stanford, California 94305, USA. ¹¹Renal Cancer Program, Dana-Farber/Harvard Cancer Center, Boston, Massachusetts 02115, USA. ¹²Department of Medicine, Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. ¹³Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹⁴Department of Dermatology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ¹⁵Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. ¹⁶Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. ¹⁷Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.

*These authors contributed equally to this work.

Fig. 2). Loss of *SMAD4* in advanced PCA is further supported by recent report of frequent epigenetic silencing of the *SMAD4* promoter in advanced disease¹². On the functional level, *SMAD4* knockdown in PC3 showed significantly enhanced frequency of metastases to the lung from renal capsule implantation (Fig. 1d and Supplementary Fig. 3). These observations prompted speculation that a *SMAD4*-dependent barrier constrains PCA progression.

To obtain genetic evidence that *Smad4* extinction enables progression, we engineered mice harbouring *Pb-Cre4* and conditional knockout alleles of *Pten* and/or *Smad4* (designated *Pten*^{pc-/-} and *Smad4*^{pc-/-}) and confirmed prostate-specific deletion (Supplementary Fig. 4). At 7 weeks of age, both *Pten*^{pc-/-} and *Pten*^{pc-/-}*Smad4*^{pc-/-} models develop low-grade PIN (Fig. 2a). Consistent with previous studies^{7,8}, *Pten*^{pc-/-} mice acquired invasive features after 19 weeks of age and most survived beyond 1 year of age (Fig. 2b). In contrast, *Pten*^{pc-/-}*Smad4*^{pc-/-} mice developed focally invasive PCA by 11 weeks (Fig. 2a, arrow) and highly aggressive invasive PCA with stromal reaction by 15 weeks of age (Fig. 2a and Supplementary Fig. 5). All *Pten*^{pc-/-}*Smad4*^{pc-/-} mice died by 32 weeks of age due largely to bladder outlet obstruction which caused hydronephrosis and renal failure (Fig. 2b, c and Supplementary Fig. 6), whereas *Smad4*^{pc-/-} mice showed no prostate neoplasia beyond 2 years of age (Fig. 2b and Supplementary Fig. 7).

Molecular pathological analysis of PCA-bearing *Pten*^{pc-/-}*Smad4*^{pc-/-} mice showed metastatic spread of Krt8 and androgen receptor-positive (Krt8⁺, Ar⁺) tumour nodules to draining lumbar lymph nodes in 25/25 cases and lung metastases in 3/25 cases (0.3–3 mm diameter metastatic nodules) (Fig. 2d, Supplementary Fig. 8 and Supplementary Table 1). The histological features of these metastases resembled those of the primary prostate tumour (Fig. 2d). These observations are in contrast to the *Pten*^{pc-/-} PCA-bearing mice which never developed metastatic lesions when examined at 1 year of age (*n* = 10), and only two mice (2/8) older than 1.5 years of age contained a solitary lumbar lymph node metastasis and one of these mice also possessed a solitary lung micrometastasis (Supplementary Table 1), a constrained progression

phenotype that aligns with previous reports^{7–9}. Similarly, 0/20 *Pten*^{pc-/-}*p53*^{pc-/-} PCA-bearing mice developed metastasis during the same observation period (data not shown).

Having demonstrated the distinctly different metastatic potential of the *Pten*^{pc-/-}, *Pten*^{pc-/-}*Smad4*^{pc-/-}, and *Pten*^{pc-/-}*p53*^{pc-/-} models, we then compared transcriptomes of primary PCAs from each to gain insight into the molecular determinants of their phenotypic differences. First, primary anterior prostate tumours with comparable sizes were harvested from 15-week-old animals from each model for mRNA profiling. Comparisons of *Pten*^{pc-/-}*Smad4*^{pc-/-} (*n* = 5) versus *Pten*^{pc-/-} (*n* = 5) or *Pten*^{pc-/-}*p53*^{pc-/-} (*n* = 3) with *Pten*^{pc-/-} (*n* = 5) prostate tumour transcriptomes defined the *Pten*^{pc-/-}*Smad4*^{pc-/-} or *Pten*^{pc-/-}*p53*^{pc-/-} signatures (Supplementary Data 2, 3). Ingenuity Pathway Analysis (IPA) was used to generate hypotheses on the biological processes that underlie the metastatic phenotype in the *Pten*^{pc-/-}*Smad4*^{pc-/-} PCAs. In contrast to the *Pten*^{pc-/-}*p53*^{pc-/-} signatures, we found that the two most significantly enriched gene-categories in the *Pten*^{pc-/-}*Smad4*^{pc-/-} signature are 'cellular movement' and 'cellular growth and proliferation' (Supplementary Fig. 9).

Enrichment of cell growth and proliferation genes in *Pten*^{pc-/-}*Smad4*^{pc-/-} PCA concurs with histopathological observations of markedly increased proliferation index relative to *Pten*^{pc-/-} tumours (Fig. 3a, b). Increased proliferation index was not associated with changes in apoptosis (Supplementary Fig. 10), but rather neutralization of oncogene-induced senescence (OIS) as reflected by loss of senescence-associated β -galactosidase staining (Fig. 3a, b). A survey of key regulators of G1/S transition and OIS revealed significant induction of cyclin D1 protein but without significant changes in p53, p21 (also known as Cdkn1a) and p27 (also known as Cdkn1b) in *Pten*^{pc-/-}*Smad4*^{pc-/-} relative to *Pten*^{pc-/-} tumours (Fig. 3c and Supplementary Fig. 11). Complementing this hypothesis-driven survey, cyclin D1 was computationally identified as the only cell cycle regulator in the *Pten*^{pc-/-}*Smad4*^{pc-/-} signature that both exhibits human PCA progression-correlated expression in Oncomine and harbours putative SMAD-binding elements (SBEs) in its promoter

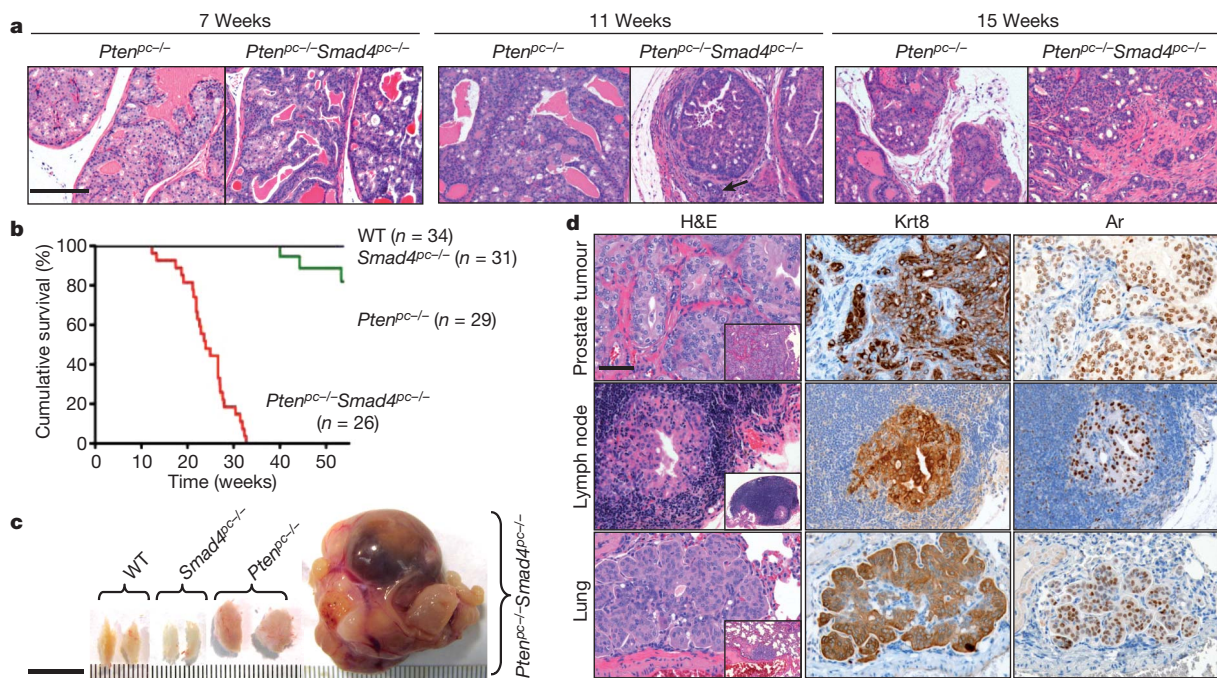


Figure 2 | *Smad4* deletion drives progression of *Pten*-deficient prostate tumour to highly aggressive prostate cancer metastatic to lymph node and lung. **a**, Haematoxylin and eosin (H&E) stained sections of representative anterior prostates (AP) at 7, 11 and 15 weeks. Scale bar, 200 μ m. **b**, Kaplan–Meier cumulative survival analysis showing significant (*P* < 0.0001) decrease in

lifespan in the *Pten*^{pc-/-}*Smad4*^{pc-/-} compared with the *Pten*^{pc-/-} cohort. **c**, Gross anatomy of representative prostates at 22 weeks of age. Scale bar, 10 mm. **d**, H&E-stained sections and immunohistochemical analyses of primary PCA, lumbar lymph nodes and lung of *Pten*^{pc-/-}*Smad4*^{pc-/-}. The tumour context is depicted in low-magnification insets. Scale bar, 50 μ m.

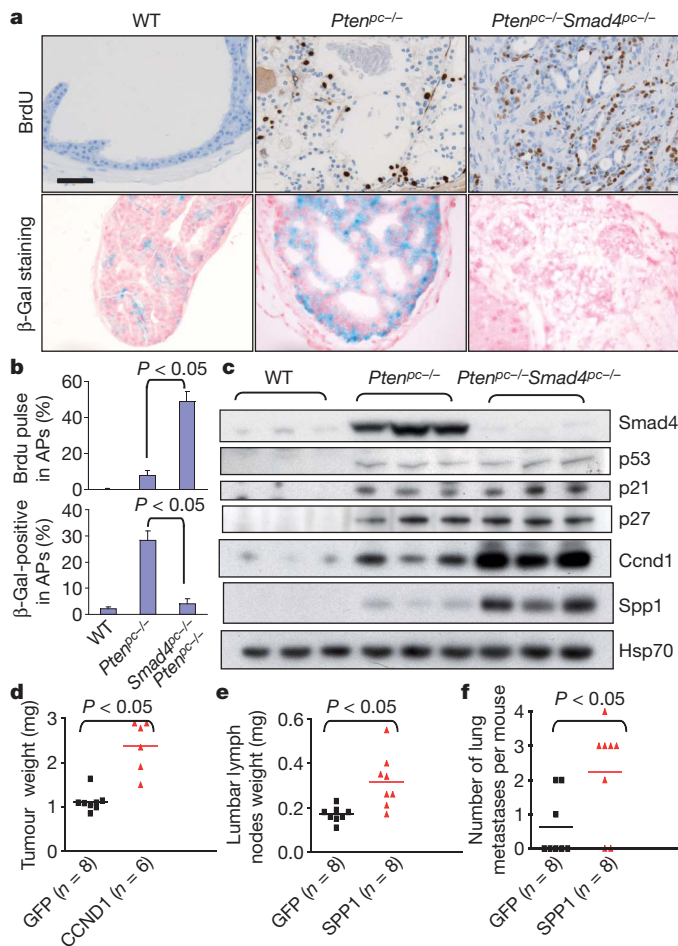


Figure 3 | Ccnd1 and Spp1 are mediators of prostate tumour cell proliferation and metastasis. **a**, BrdU pulse-labelling and SA-β-galactosidase (β-Gal) staining of 15-week-old APs. **b**, Quantification of BrdU pulse labelling and β-Gal staining. Error bars represent s.d. for a representative experiment performed in triplicate. **c**, Western blot analysis demonstrating elevated Ccnd1 and Spp1 levels in *Pten^{pc/pc}-Smad4^{pc/pc}* compared to *Pten^{pc/pc}* prostate tumours. **d**, Enforced CCND1 expression significantly enhanced prostate xenograft tumour growth of PC3 cells. **e**, **f**, Enforced SPP1 expression significantly increases metastatic activity of PC3 cells from prostate xenograft to lumbar lymph nodes (**e**) and to lung (**f**).

(Supplementary Data 2). Indeed, chromatin immunoprecipitation (ChIP) assays confirmed that SMAD4 can bind to one of the SBEs in the cyclin D1 gene promoter (Supplementary Figs 12 and 13). Correspondingly, TGFβ1 (also known as TGFB1)-treated SMAD4-transduced *Pten^{pc/pc}-Smad4^{pc/pc}* prostate tumour cells show down-regulated cyclin D1 expression (Supplementary Fig. 14a). Finally, enforced cyclin D1 expression significantly enhanced xenograft tumour growth *in vivo* (Fig. 3d). Together, these data support the thesis that cyclin D1 is a key mediator of the cardinal tumour biological feature of increased proliferation in the metastatic *Pten^{pc/pc}-Smad4^{pc/pc}* model.

We next obtained available ORFs corresponding to 21 of the 84 'Cellular Movement' genes (Supplementary Table 2) and assayed their ability to enhance invasion of human prostate cancer cells. Using the modified Boyden chamber assay, 10/21 ORFs enhanced invasion of prostate cancer cells including PC3 (Supplementary Table 2). Among these validated invasion genes, *SPP1* was selected for deeper analysis given its PCA progression-correlated expression in Oncomine, its prognostic potential for BCR in univariate COX proportional hazard analysis in a data set comprising of transcriptome and outcome data on 79 PCA patients (Supplementary Tables 3 and 4)¹³, and its known link to TGFβ signalling under different cellular contexts^{14–16}. Western blotting and immunohistochemical analyses confirmed increased Spp1

expression in *Pten^{pc/pc}-Smad4^{pc/pc}* compared to *Pten^{pc/pc}* tumours (Fig. 3c and Supplementary Fig. 11) and promoter analysis¹⁷ identified a conserved SBE in the *Spp1* promoter which was confirmed by ChIP assay in cells treated with TGFβ1 (Supplementary Fig. 15). In contrast to previous studies showing SMAD4 as an inducer of *Spp1* expression through displacement of transcription repressors from *Spp1* promoter in a mink lung epithelial cell line and a pre-osteoblastic cell line^{14,16}, loss of *Smad4* in the *Pten^{pc/pc}-Smad4^{pc/pc}* prostate tumour cells results in markedly increased Spp1 expression (Fig. 3c and Supplementary Data 2). TGFβ1 treatment correspondingly suppressed *Spp1* expression in SMAD4-dependent manner in *Pten^{pc/pc}-Smad4^{pc/pc}* prostate tumour cells (Supplementary Fig. 14b). These observations underscore the context-specific actions of TGFβ-SMAD4 signalling on its downstream targets¹⁸. Next, to verify that Spp1 functionally contributes to the metastatic phenotype in our model, we showed significant inhibition of invasive activity *in vitro* upon knockdown of *Spp1* in *Pten^{pc/pc}-Smad4^{pc/pc}* mouse PCA cells (Supplementary Fig. 16). Conversely, enforced SPP1 expression enhanced invasion *in vitro* of several human lines (Supplementary Fig. 17). Finally, orthotopic implantation of SPP1-transduced PC3 cells in the prostate exhibited increased lumbar lymph node metastasis and enhanced metastasis to lung (Fig. 3e–f and Supplementary Fig. 18). These results strongly indicated that *SPP1* is a pro-metastasis invasion gene in human PCA and in the *Pten^{pc/pc}-Smad4^{pc/pc}* PCA model.

The *in vivo* genetic modelling studies, the *in silico* transcriptomic and pathway analyses, along with the tumour biological and functional characterizations collectively point to the inactivation of *Pten* and *Smad4* as well as activation of cyclin D1 (also known as Ccnd1) and Spp1 as drivers of PCA progression. As such, we posited that these four key PCA metastasis progression relevant genes may carry prognostic value for metastasis risk in human PCA (see Supplementary Fig. 19). To this end, we assessed how robustly these four genes can stratify risk of BCR (>0.2 ng ml⁻¹) in the data set from ref. 13. Although only *SPP1* was significantly correlated with BCR in univariate analysis, an overall risk score integrating the four-gene signature by multivariate Cox regression showed significant association with BCR as well (P -value = 0.0025, and overall C-index = 0.66, see Supplementary Tables 4 and 5). Furthermore, the four-gene model robustly stratified the ref. 13 cohort by *K*-mean clustering into two groups that exhibited significant difference in risk for BCR by Kaplan–Meier analysis (Fig. 4a; hazard ratio = 2.6, log-rank test P = 0.012). Importantly, by *C*-statistics, this four-gene signature carries independent prognostic information as it can enhance the prognostic accuracy of Gleason score from C-index from 0.77 to 0.8 (Fig. 4b), even though by itself, the four-gene signature (C-index as 0.75) performs only as well as Gleason score alone (Fig. 4b).

We repeated this analysis in an independent extreme-case-control cohort derived from the Physicians' Health Study (PHS) (Supplementary Table 6; see Methods for study design), where we showed that the four-gene model was also capable of enhancing the prognostic accuracy of Gleason score in predicting metastatic lethal outcome (Fig. 4c; C = 0.716 by four-gene signature). Although exclusion of non-informative cases may have biased towards a positive association, the prognostic performance by this four-gene signature is unlikely a chance occurrence because, by gene-set-enrichment testing, it outperforms 243 other bidirectional signatures curated in the Molecular Signature Databases of the Broad Institute (MSigDB, version 2.5) in predicting metastatic lethal outcome in this PHS extreme-case-control cohort (Supplementary Fig. 20).

Encouraged by the prognostic value in two independent cohorts using RNA expression yet mindful of the inherent intra-tumoural heterogeneity of PCA which may obscure expression differences in whole-tumour transcriptome profiles, we next performed immunohistochemical staining with validated antibodies against PTEN, SMAD4, cyclin D1 and SPP1 on a tumour tissue microarrays (TMA) comprising a cohort of 405 tumour specimens randomly selected from men diagnosed with prostate cancer who underwent

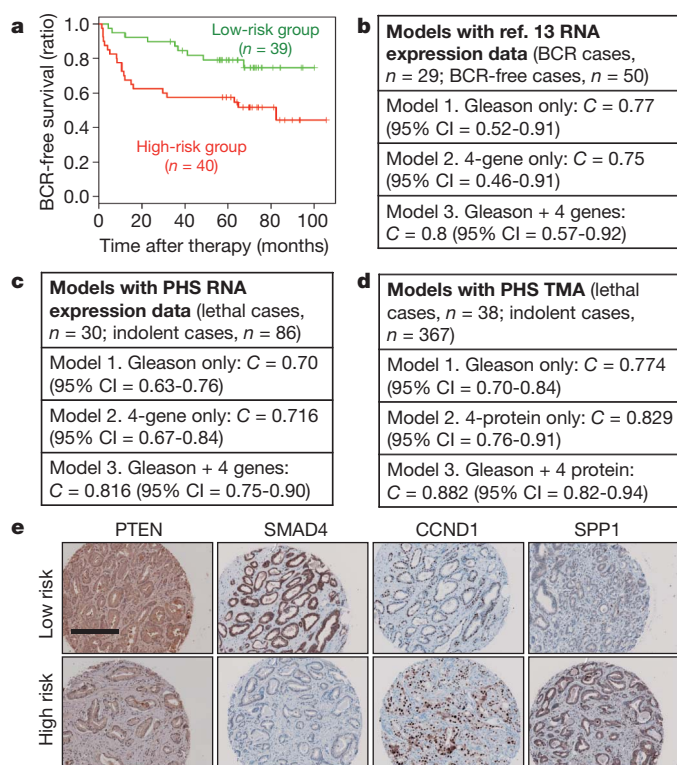


Figure 4 | Prognostic potential of a four-gene signature in human PCA. **a**, The four-gene set of *PTEN/SMAD4/CCND1/SPP1* can dichotomize PCA cases for BCR in the ref. 13 data set. **b**, **c**, C-statistic analysis revealed that this four-gene set can enhance the prognostic accuracy of Gleason score in the ref. 13 data set (**b**) and in an independent PHS cohort (**c**). **d**, TMA-based four-protein model also significantly improve the prognostic ability of Gleason ($P = 0.015$) from the PHS cohort. **e**, Representative immunohistochemical staining with specific antibody against PTEN, SMAD4, CCND1 and SPP1 in the Directors Challenge TMA. Scale bar, 200 µm.

radical prostatectomy in the PHS cohort. Staining results were quantified by expert pathologists (R.L. and M.L.) blinded to the outcome of the cases. Indeed, not only does the four-protein model improve the prognostic accuracy of Gleason score in combination, it performs significantly better than Gleason score alone (Fig. 4d; $C = 0.774$ for Gleason only, $C = 0.829$ for four-protein model alone, and $C = 0.882$ for Gleason + four-protein model; $P = 0.015$ for improvement). Moreover, the addition of the four-protein model to the clinical parameters (Gleason, age at diagnosis, TNM stage; $C = 0.842$) leads to a significant seven point increase in the C-statistic ($C = 0.913$), P -value for difference between full clinical model versus clinical model + four-protein signature = 0.047 (Supplementary Table 7). The enhanced prognostic value of 'Gleason + four-protein model' was similarly validated in yet another independent cohort, the Directors Challenge TMA containing 40 prostate cancer patients with recurrence as outcome (Supplementary Table 8) (Fig. 4e and Supplementary Fig. 19c; $C = 0.704$ for Gleason alone versus $C = 0.740$ for Gleason + four-protein model).

In summary, concomitant *Pten* and *Smad4* inactivation in the prostate epithelium can bypass OIS, enhance tumour cell proliferation and drive invasion to produce a fully-penetrant invasive and metastatic PCA phenotype in the mouse (Supplementary Fig. 21). The human relevance of this *Pten*^{pc/-}*Smad4*^{pc/-} model of metastatic PCA is credentialed by the prognostic significance of a four-marker signature derived from this mouse model in predicting biochemical recurrence or lethal metastasis in human PCAs. Thus this study will facilitate the development of a molecularly-based prognostic assay that may complement the current standard of care to improve evidence-based management of PCA patients, a current major unmet need.

METHODS SUMMARY

All mice were maintained in pathogen-free facilities under institutionally approved protocols. At the time of sacrifice, tissues were collected and processed as described in Methods. For microarray analysis, prostate tissues from wild-type, *Pten*^{pc/-}, *Pten*^{pc/-}*Smad4*^{pc/-} and *Pten*^{pc/-}*p53*^{pc/-} mice were isolated and total mRNA extracted, labelled and hybridized to Affymetrix Mouse Genome 430 2.0 chips and resultant transcriptomes were compared to generate phenotype-based differentially expressed gene lists. The Ingenuity Pathways Analysis program was used to analyse further the enrichment of canonical pathways, molecular and cellular functions. Validation assays for invasion were performed in standard 24-well invasion chambers containing Matrigel. Based on both hypothesis-driven and computational approaches, we developed a four-marker signature comprising of PTEN, cyclin D1, SMAD4 and SPP1. The four-gene markers (mRNA levels) were assayed on two independent mRNA expression data set and the corresponding four-protein markers (antibody assays by IHC) were assayed on two independent TMAs (see Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 November 2009; accepted 16 November 2010.

Published online 2 February 2011.

- Andreou, M. & Cheng, L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. *Hum. Pathol.* **41**, 781–793 (2010).
- Jemal, A., Siegel, R., Xu, J. & Ward, E. Cancer statistics, 2010. *CA Cancer J. Clin.* (2010).
- Walsh, P. C., DeWeese, T. L. & Eisenberger, M. A. Localized prostate cancer. *N. Engl. J. Med.* **357**, 2696–2705 (2007).
- Rubin, M. A. Targeted therapy of cancer: new roles for pathologists—prostate cancer. *Mod. Pathol.* **21** (Suppl 2), S44–S55 (2008).
- Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
- Jeet, V., Russell, P. J. & Khatri, A. Modeling prostate cancer: a perspective on transgenic mouse models. *Cancer Metastasis Rev.* **29**, 123–142 (2010).
- Wang, S. *et al.* Prostate-specific deletion of the murine *Pten* tumor suppressor gene leads to metastatic prostate cancer. *Cancer Cell* **4**, 209–221 (2003).
- Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of *Pten*-deficient tumorigenesis. *Nature* **436**, 725–730 (2005).
- Trotman, L. C. *et al.* *Pten* dose dictates cancer progression in the prostate. *PLoS Biol.* **1**, E59 (2003).
- Ma, X. *et al.* Targeted biallelic inactivation of *Pten* in the mouse prostate leads to prostate cancer accompanied by increased epithelial cell proliferation but not by reduced apoptosis. *Cancer Res.* **65**, 5730–5739 (2005).
- Zeng, L., Rowland, R. G., Lele, S. M. & Kyprianou, N. Apoptosis incidence and protein expression of p53, TGF- β receptor II, p27Kip1, and Smad4 in benign, premalignant, and malignant human prostate. *Hum. Pathol.* **35**, 290–297 (2004).
- Aitchison, A. A. *et al.* Promoter methylation correlates with reduced Smad4 expression in advanced prostate cancer. *Prostate* **68**, 661–674 (2008).
- Glinsky, G. V., Glinskii, A. B., Stephenson, A. J., Hoffman, R. M. & Gerald, W. L. Gene expression profiling predicts clinical outcome of prostate cancer. *J. Clin. Invest.* **113**, 913–923 (2004).
- Hullinger, T. G., Pan, Q., Viswanathan, H. L. & Somerman, M. J. TGF β and BMP-2 activation of the OPN promoter: roles of Smad- and Hox-binding elements. *Exp. Cell Res.* **262**, 69–74 (2001).
- Packer, L. *et al.* Osteopontin is a downstream effector of the PI3-kinase pathway in melanomas that is inversely correlated with functional PTEN. *Carcinogenesis* **27**, 1778–1786 (2006).
- Shi, X., Bai, S., Li, L. & Cao, X. Hoxa-9 represses transforming growth factor- β -induced osteopontin gene transcription. *J. Biol. Chem.* **276**, 850–855 (2001).
- Paik, J. H. *et al.* FoxOs cooperatively regulate diverse pathways governing neural stem cell homeostasis. *Cell Stem Cell* **5**, 540–553 (2009).
- Massague, J., Seoane, J. & Wotton, D. Smad transcription factors. *Genes Dev.* **19**, 2783–2810 (2005).
- Yu, Y. P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.* **22**, 2790–2799 (2004).
- Dhanasekaran, S. M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors are grateful to the late W. Gerald for providing the primary gene expression data and clinical outcome files¹³. We thank S. Zhou for excellent mouse husbandry and care, B. Xiong and G. Tonon for bioinformatic assistance, and S. Jia, J. M. Stommel, J. Paik, M. Kim and A. C. Kimmelman for helpful discussion. We thank M. Vidal, the Ellison Foundation and DFCI ISR for support of ORF cloning efforts, R. Maser for MSCV-puro-v5 gateway vector, W. Hahn for shRNA constructs. We thank the DF/HCC Specialized Histopathology Core and the DF/HCC Tissue Microarray and Imaging core for the TMA IHC staining; the DFCI/BWH Center for Molecular Oncologic Pathology (CMOP) for the quantification of the IHC. Z.D. was supported by the Damon Runyon Cancer Research Foundation. D.H. was supported by

a graduate fellowship from the National Science Foundation. H.Z. was supported by the Helen Hay Whitney Foundation. Y.A.W. was supported by the Multiple Myeloma Research Foundation. This work is supported by the Belfer Institute for Applied Cancer Science, NCI U01-CA84313 (L.C. and R.A.D.), DF/HCC SPORE in Prostate Cancer P50 CA090381-08 (Z.D.), the National Cancer Institute (M.L. R01CA131945 and P50 CA90381, L.M. R01 5R01CA136578, M.S. R01CA141298), and the Linda and Arthur Gelb Center for Translational Research (M.L.). R.A.D. was supported by an American Cancer Society Research Professorship and L.M. was supported by the Prostate Cancer Foundation.

Author Contributions Z.D. designed and performed the experiments. L.C. and R.A.D. supervised experiments and computational analysis and contributed as senior authors. C.J.W., Y.X., Y.H., D.H., T.R.G., M.J.S., W.H.W. and L.M. performed the

computational analysis. G.C.C. provided pathology analyses. X.W., R.L., S.S. and M.L. performed TMA staining and quantification. N.B. generated *Smad4L* mouse allele. D.E.H. provided the human ORFeome clones. D.H., J.Z., S.R.P., E.S.L., B.H., S.J., H.Z., A.H.S. and K.L.S. performed the experiments. Y.A.W. contributed to the writing of the manuscript.

Author Information The microarray data have been deposited in the GEO database with accession number GSE25140. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to ron_depinho@dfci.harvard.edu or lynda_chin@dfci.harvard.edu

METHODS

***Pten* and *Smad4* conditional alleles, genotyping and expression analysis.** The *Pten*^{loxP} and *Smad4*^{loxP} conditional knockout alleles have been described elsewhere^{21,22}. *p53*^{loxP} strain was generously provided by A. Berns²³. Prostate epithelium-specific deletion was effected by the PB-Cre4²⁴ and was obtained from MMHCC (http://mouse.ncicrf.gov/search_results.asp). All cohorts were in a FVB/n, C57BL/6 and 129/Sv mixed genetic background.

Tissue analysis. Normal and tumour tissues were fixed in 10% neutral-buffered formalin overnight then processed, paraffin-embedded, sectioned and stained with haematoxylin and eosin according to standard protocol. For immunohistochemistry, 5 µm sections were incubated with primary antibodies overnight at 4 °C in a humidified chamber. Primary antibodies: rabbit polyclonal anti-androgen receptor (06-680, Millipore), Smad4 (1676-1, Epitomics), Ck8 (also known as Krt8) (GTX15465, GeneTex); p53 (VP-P956, Vector Laboratories), p21 (C-19, sc-397, Santa Cruz), p27 (2747-1, Epitomics) and Cyclin D1 (RM-9104-R7, Thermo Scientific); and mouse monoclonal Spp1 (sc-21742, Santa Cruz). For rabbit antibodies, sections were subsequently developed using Dako Envision. Mouse monoclonal staining was developed using MOM kit (Vector). To assay senescence in prostate tissue of the various genotypes, frozen sections were stained for SA-β-Gal as described elsewhere⁷. Representative sections from at least three mice were counted for each genotype.

For western blot analysis, tissues and cells were lysed in RIPA buffer (20 mM Tris pH 7.5, 150 mM sodium chloride, 1% Nonidet P-40, 0.5% sodium deoxycholate, 1 mM EDTA, 0.1% SDS) containing complete mini protease inhibitors (Roche) and phosphatase inhibitors. Western blots were obtained using 20–50 µg of lysate protein, and were incubated with antibodies against Smad4 (sc-7966, Santa Cruz), phospho-Akt^{Ser473} (4060, Cell Signaling Technology), Akt (3272, Cell Signaling Technology), V5 (R960-25, Invitrogen), Hsp70 (610607, BD Transduction Laboratories), and Spp1 (sc-21742, Santa Cruz), p53 (sc-6243, Santa Cruz), p27 (2747-1, Epitomics), p21 (65961A, BD Biosciences), Cyclin D1 (2926, Cell Signaling), pSmad1/5/8 (9511, Cell Signaling), Smad1 (9743, Cell Signaling), pSmad2 (Ser465/467) (3101S, Cell Signaling), Smad2 (3103, Cell Signaling), pSmad3 (ab52903, Abcam), Smad3 (06-920, Millipore).

Establishment of mouse prostate tumour cell lines. Tumours were dissected from prostates of *Pten*^{loxP/loxP}*Smad4*^{loxP/loxP} PB-Cre4⁺ (*Pten*^{pc-/-}*Smad4*^{pc-/-}) mice, minced, and digested with 0.5% type I collagenase (Invitrogen) as described previously. After filtering through a 40-µm mesh, the trapped fragments were plated in tissue culture dishes coated with type I collagen (BD Pharmingen). Cells with typical epithelial morphology were collected, and single cells were seeded into each well of a 96-well plate. Three independent cell lines (*Pten*^{pc-/-}*Smad4*^{pc-/-}-1, -2 and -3,) were established and maintained in DMEM plus 10% fetal bovine serum (FBS, Omega Scientific), 25 µg ml⁻¹ bovine pituitary extract, 5 µg ml⁻¹ bovine insulin, and 6 ng ml⁻¹ recombinant human epidermal growth factor (Sigma-Aldrich). The prostate tumour epithelial cells express epithelial marker CK8 detected by immunofluorescence analyses using CK8 (GTX15465, GeneTex) antibody.

Establishment of inducible *Pten*^{pc-/-}*Smad4*^{pc-/-} SMAD4-TetOn cell lines. *Pten*^{pc-/-}*Smad4*^{pc-/-} prostate tumour cells (see above) were used as parental cells for establishment of inducible SMAD4 TetOn cells using TetOn Advanced Inducible Gene Expression System (Clontech). Human SMAD4 coding region inserted into the pTRE-Tight vector, and a TetOn SMAD4 expression system was generated according to the manufacturer's protocol. Stable clones were induced to express SMAD4 using 1 µg ml⁻¹ doxycycline (dox), and expression was verified to be comparable to the SMAD4 level in *Pten*^{pc-/-} prostate tumours by western blot analysis of whole-cell extracts, using anti-SMAD4 antibody (sc-21742, Santa Cruz) (Supplementary Fig. 12).

RNA isolation and real-time PCR. Total RNA was extracted using TRIzol followed by RNeasy Mini kit (Qiagen) cleanup and RQ1 RNase-free DNase Set treatment (Promega) according to the manufacturer's instructions. First strand cDNA was synthesized using 1 µg of total RNA and Superscript II (Invitrogen). Real-time quantitative PCR was performed in triplicates with a MxPro3000 and SYBR GreenER qPCR mix (Invitrogen). The relative amount of specific mRNA was normalized to *Gapdh*. Primer sequences are available upon request.

Transcriptomic and pathway analyses. For transcriptomic analyses, anterior prostate from mice at 15 weeks of age were isolated and total mRNA extracted, labelled and hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 Arrays by the Dana-Farber Cancer Institute Microarray Core Facility according to the manufacturer's protocol. Affymetrix mouse MOE430 raw data (CEL files) were pre-processed using robust multi-array analysis (RMA) of the Affy package of Bioconductor. The background-corrected, normalized and summarized probe set intensity data were then analysed using significance analysis of microarrays (SAM) to identify differentially expressed genes. Using a twofold, FDR 5% cut-off, we generated a 3,532 probe set that distinguishes differentially expressed genes in

anterior prostate samples from *Pten*^{pc-/-} (five mice) versus WT (PB-Cre4) (three mice), 397 probe sets that distinguishes differentially expressed genes in anterior prostate samples from *Pten*^{pc-/-}*Smad4*^{pc-/-} (five mice) versus *Pten*^{pc-/-} (five mice), and 370 probe sets that distinguishes differentially expressed genes in *Pten*^{pc-/-}*p53*^{pc-/-} (three mice) versus *Pten*^{pc-/-} (five mice). Gene information for all probes was annotated based on 'Mouse430_2.na28.annot.csv' downloaded from the Affymetrix website. Probes with multiple genes in the Affymetrix annotation file were mapped against latest mouse genome build (UCSC mm9) for the single matching gene. Probes mapped to more than one position on mm9 were ignored. Human orthologues of mouse genes were extracted from HomoloGene build 64 (<ftp://ftp.ncbi.nih.gov/pub/HomoloGene/>). Intersection of the murine list with the human orthologous genes produced an orthologous set of genes.

All differentially expressed gene lists generated as described above were further analysed with the Ingenuity Pathways Analysis program (<http://www.ingenuity.com/index.html>) to identify canonical pathways, and molecular and cellular functions enriched in the related gene lists.

cDNA and shRNA constructs. Human cDNAs presented in Supplementary Table 1 were obtained from the Human ORFeome collection, Japan National Institute of Technology and Evaluation (NITE), Japan, and transferred into a modified pMSCV-V5 vector via Gateway recombination. Knockdown of human SMAD4 and mouse Spp1 were performed by infecting the indicated cells with lentivirus containing either *shSMAD4* or *shSpp1* (provided by W. Hahn). The shRNA constructs for *shSMAD4* #1, #2 correspond to clone IDs# TRCN0000040028 (hairpin sequence: CCGGGCAGACAGAACTGGATTAACTCGAGTTTAAATCCAGT TTCTGTCTGCTTTTGTG), and TRCN0000040029 (hairpin sequence: CCGGCC TGAGTATTGGTGTTCATTCTCGAGAATGGAACACCAATACTCAGGTT TTTG), respectively. The shRNA constructs for *shSpp1* #1, #2 correspond to clone IDs# TRCN0000054698 (Hairpin sequence: CCGGCTCTTAGCTTA GTCTGTTGTTCTCGAGAACAAACAGACTAAGCTAAGAGTTTGTG), and TRCN0000054700 (Hairpin sequence: CCGGCACAAGGACAAGCTAGTCC TACTCGAGTAGGACTAGCTGTCTCTGTGTTTGTG), respectively, in the RNAi Consortium (TRC).

Viral production and transduction. Approximately 2×10^6 293T cells were seeded in 100 mm plates 15 h before transfection (~30% confluent) in 10% FBS/DMEM with antibiotics. For MSCV viral production, 3 µg viral backbone, 2.7 µg gag/pol expression vectors, and 0.3 µg VSV-G expression vector were diluted to 20 µl using Opti-MEM (Invitrogen) and combined with 180 µl Opti-MEM containing 12 µl FuGENE-6 (Roche). This mixture was incubated at room temperature (RT) for 20 min and added to the 10 ml media covering the 293T cells. For pLKO shRNA lentivirus production, 10 µg of viral backbone and 10 µg of lentiviral packaging vectors were diluted to 1,000 µl using Opti-MEM (Invitrogen). The resulting mix was combined with 1,000 µl Opti-MEM containing 30 µl Lipofectamine2000 (Invitrogen), incubated at room temperature for 20 min and added to 8 ml media covering the 293T cells. The media was replaced with 10% FBS/DMEM approximately 10 h post-transfection and viral supernatants were collected at 36 h and 60 h after transfection and combined. Viral supernatants (5 ml) containing 8 µg ml⁻¹ polybrene were added to target cells that were seeded 24 h before infection at 70–80% confluence. Cells were infected twice and allowed to recover in 10% FBS/RPMI 1640 with antibiotics for 12 h following the second infection, after which cells were selected with 2 µg ml⁻¹ puromycin for 4 days and allowed to recover in normal medium for 24 h before further experiments.

Transwell invasion assay. Standard 24-well Boyden invasion chambers (BD Biosciences) were used to assess cell invasiveness following the manufacturer's suggestions. Briefly, cells were trypsinized, rinsed twice with PBS, resuspended in serum-free media, and seeded at 2×10^5 cells per well for PC3 cells and *Pten*^{pc-/-}*Smad4*^{pc-/-} cells, 4×10^5 cells per well for BPH1 cells. Chambers in triplicate were placed in 10% serum-containing media as a chemo-attractant and an equal number of cells were seeded in cell culture plates in triplicate as input controls. Following 22 h incubation, chambers were fixed in 10% formalin, stained with crystal violet for manual counting or by pixel quantification with Adobe Photoshop. Data was normalized to input cells to control for differences in cell number (loading control).

Orthotopic and renal capsule implantation. Male SCID mice (6 weeks old) were obtained from Taconic. Orthotopic and renal capsule implantations were performed as described previously^{25,26}. Briefly, a suspension of 1×10^6 cells in 50 µl of a 1:1 mixture of PBS and Matrigel (BD Biosciences) was injected into the anterior prostate lobe. For renal capsule implantation 5×10^5 cells were suspended in 50 µl of neutralized type I rat tail collagen (BD Biosciences), allowed to gel at 37 °C for 15 min, covered with growth medium, followed by grafting beneath the renal capsule of mice.

Identification of putative SMAD binding sites (SBEs). The Smad binding elements (SBEs) in the promoters of the *Pten*^{pc-/-}*Smad4*^{pc-/-} signature of 267 genes

were identified computationally by established methods¹⁶. Briefly, the conserved nucleotides in the 4kb promoter regions of the promoters were isolated and scanned for enrichment of the SMAD binding motifs in TRANSFAC. Enrichment was assessed by comparing the target regions to matched control regions at the same distance from the transcription start sites of random genes. Promoter analysis on these gene sets for SBEs used the CisGenome software (<http://www.biostat.jhsph.edu/~hji/cisgenome/>).

Chromatin immunoprecipitation (ChIP) assay. ChIP assays with 1 µg of normal mouse IgG (Upstate), normal rabbit IgG (Upstate), anti-RNA polymerase II (PolII) (Upstate), anti-acetyl-Histone H3 (Upstate) or anti-SMAD4 IgG (mouse monoclonal, clone B8, sc-7966, Santa Cruz) overnight at 4 °C were conducted by established methods¹⁶.

Immunohistochemical evaluation of outcome tissue microarrays (TMAs). Immunohistochemical staining was performed on 5-µm sections of the TMAs to assess cytoplasmic PTEN (PN37, rabbit polyclonal, 18-0256, Zymed), cytoplasmic SMAD4 (mouse monoclonal, clone B8, sc-7966, Santa Cruz), nuclear cyclin D1 (Rabbit monoclonal, SP4, RM-9104-R7, Thermo Scientific), and cytoplasmic SPP1 expression (Rabbit polyclonal, O17, 18625, IBL) after citrate-based antigen retrieval.

TMA slides were scanned using the CRi Nuance v2.8 (Woburn) slide scanner following the standard bright field TMA protocol. The system acquires images at 20 nm intervals and combines them into a stack file which represents one image. This was done automatically to create one image for each core on the TMA. The maximum likelihood method was used to extract the spectra of DAB and haematoxylin, which represent the different elements of IHC. inForm v0.4.2 software (CRi) was used to analyse the spectral images of each core. Initially, a training set comprising two classes of tissue was created: 'tumor' and 'other'. Representative areas for each class were marked on 12–16 images from each TMA. The software was trained on these areas using the spectra of both the counterstain (haematoxylin) and the immunostain (DAB) and tested to determine how accurate it could differentiate between the two classes. This process was repeated until further iterations no longer improved accuracy.

Histological images were then analysed using the 'nuclear or cytoplasmic' algorithm. The multispectral imaging capabilities of the Nuance slide scanner allows the software to isolate or segment the nuclei using the unmixed spectra of the nuclear counterstain and the DAB immunohistochemical stain used in addition for a nuclear biomarker. In turn, cytoplasm is found based on the non-nuclear tumour area. Threshold settings approximated: scale 1, offset subtraction 0, minimum blob size 30, maximum blob size 10,000, circularity threshold 0, edge sharpness 0, fill hole enabled (nuclear parameters); algorithm 4, area 200, compactness 0.5, Wht threshold 225 (cytoplasmic parameters). The final score was based on the percentage of the cytoplasmic or nuclear tumour area that was positively stained and this was represented as a ten bin histogram. This involved each pixel being placed into one of ten bins based on the intensity of the DAB spectra, with an adjustment of the threshold for the 9th bin by the user in order to create a desirable distribution. By reviewing images and their scores, a threshold level of these bins was determined that represented real staining, and the values from the bins above this threshold were added together to create a final score which represented the percentage of cytoplasmic or nuclear area that was positively stained. All samples were also reviewed by pathologists (R.L. and M.L.) to ensure that assigned scores were appropriate. TMA cores that were difficult to classify (due to technical artefacts such as folds in the tissue, air bubbles, cores overlapping or due to difficulty in morphological classification) were either eliminated from the analysis in order to categorize the tissue appropriately. The Directors Challenge TMA originally contained 52 patient samples²⁷. However, as is typical of most heavily used TMAs, some of the samples become exhausted over time from extensive use by the M.L. lab and the community. After careful quality control of each core on the TMA by R.L. in M.L. lab, only 40 high quality core samples were considered usable (Supplementary Table 7). Careful quality control of each core on the PHS TMA by R.L. in the M.L. lab, 405 high quality core samples were considered usable (Supplementary Table 5).

Clinical outcome analysis. The raw Affymetrix HG-U133A expression profiles and clinical information of 79 prostate cancer patients from the ref. 13 cohort (Supplementary Table 2)¹² were generously provided by W. Gerald. The raw data set was analysed by MAS5 algorithm. Low-expression probesets with less than 20% present calls across the 79 samples were excluded from the data. The remaining 13,027 probesets map to 8,763 genes with unique symbols, and the mean log-transformed probeset levels were used as the gene expression profiles.

A univariate Cox proportional hazard analysis was conducted using the R 'survival' package for invasion assay positive genes to identify those expression in PCA tumours was positively associated with biochemical recurrence (BCR, defined by post-op PSA > 0.2 ng ml⁻¹) in the ref. 13 data set¹².

K-means clustering algorithm was used with the *PTEN/SMAD4/CCND1/SPP1* four-gene model to identify two cancer sample clusters. The initial centres for the K-means clustering were set at the two cases with the longest Euclidean distance. Kaplan–Meier analysis for the survival difference of the two cancer patient clusters was conducted using the R 'survival' package. C-statistics analysis was conducted using the R 'survcomp' package. The statistical procedures used in the analyses include a bootstrapping step that estimates the distribution of C-statistics of all models across 10,000 random bootstrapping instances, and a comparative step that uses the paired *t*-test to compare the C-statistics of models and evaluate the statistical significance²⁸. Multivariate Cox proportional hazards model analysis with the four-gene signature was used to estimate the coefficients of individual genes, which combined the four-gene expression levels into an integrated risk score model defined.

To validate further the prognostic significance of this four-gene model, we repeated this analysis in an independent cohort derived from the Directors Challenge cohort²⁷ (Supplementary Table 7) and the Physicians' Health Study (PHS) cohort. PHS cohort (Supplementary Table 5): the men with prostate cancer included in this study were participants in the Physicians' Health Study (PHS), an ongoing randomized trial among US male physicians^{29,30}. The men were diagnosed with histologically-confirmed prostate cancer after randomization, between January 1983 and December 2004. We obtained archival formalin-fixed, paraffin-embedded tissue specimens, either radical prostatectomy (95%) or TURP (5%) and constructed tumour tissue microarrays for immunohistochemical analyses; 405 had sufficient tumour tissue available for this project. All men in the trial were followed for mortality, and cause of death was confirmed by a study endpoints committee. In addition, we retrieved medical records and questionnaire data on the men with prostate cancer to collect information on treatments, clinical characteristics, as well progression of the cancer. Through March 2010, 38 men of 405 had developed a lethal metastatic phenotype, defined by bony metastases or cancer-specific death.

We undertook gene expression profiling as part of a previous project to define molecular signatures in prostate cancer³¹ on a subset of the PHS included on the TMAs. As part of the sampling, we sought to maximize efficiency for studies of lethal prostate cancer by devising a study design that included men who either died from prostate cancer or developed metastases during follow up ('lethal prostate cancer' cases) or who survived at least 10 years after their diagnosis without any evidence of metastases (men with 'indolent prostate cancer'). We sought to include all lethal cancers, based on follow-up through March 2007, and took a random sample of indolent cancers for a total sample size of 116 cases. In this design, we exclude men with non-informative outcomes, namely those who died from other causes within 10 years of their prostate cancer diagnosis or had been followed for less than 10 years with no disease progression. The natural history of prostate cancer is quite long, with men dying of prostate cancer even 15 or more years after cancer diagnosis³². Thus, we excluded prostate cancer cases with less than 10 years follow-up to increase confidence on the outcome annotation since we are not seeking to estimate survival time. By focusing on long-follow-up cases, an extreme-case-control study design allows us to maximally identify lethal versus indolent prostate cancer. In addition, to minimize the potential that C-statistics estimation might be biased towards a higher lethal composition by such extreme-case-study-design, we have chosen a logistic regression analysis rather estimating survival analysis.

The tissue based studies were approved by the Institutional Review Boards of Harvard School of Public Health and Partners Healthcare.

We assessed the enrichment of the four-gene signature to that of 244 bidirectional signatures curated in the Molecular Signature Databases of the Broad Institute (MSigDB, version 2.5) by computing an enrichment statistic³³.

- Bardeesy, N. *et al.* *Smad4* is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreas cancer. *Genes Dev.* **20**, 3130–3146 (2006).
- Zheng, H. *et al.* p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. *Nature* **455**, 1129–1133 (2008).
- Marino, S., Vooijs, M., van der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of *Rb* in the external granular layer cells of the cerebellum. *Genes Dev.* **14**, 994–1004 (2000).
- Wu, X. *et al.* Generation of a prostate epithelial cell-specific Cre transgenic mouse model for tissue-specific gene ablation. *Mech. Dev.* **101**, 61–69 (2001).
- Berger, R. *et al.* Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells. *Cancer Res.* **64**, 8867–8875 (2004).
- Wang, Y. *et al.* A human prostatic epithelial model of hormonal carcinogenesis. *Cancer Res.* **61**, 6064–6072 (2001).
- Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209 (2002).
- Haibe-Kains, B., Desmedt, C., Sotiropoulos, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics* **24**, 2200–2208 (2008).

29. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N. Engl. J. Med.* **321**, 129–135 (1989).
30. Christen, W. G., Gaziano, J. M. & Hennekens, C. H. Design of Physicians' Health Study II—a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann. Epidemiol.* **10**, 125–134 (2000).
31. Sboner, A. *et al.* Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med. Genomics* **3**, 8 (2010).
32. Johansson, J. E. *et al.* Natural history of early, localized prostate cancer. *J. Am. Med. Assoc.* **291**, 2713–2719 (2004).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

Structures of APC/C^{Cdh1} with substrates identify Cdh1 and Apc10 as the D-box co-receptor

Paula C. A. da Fonseca^{1*}, Eric H. Kong^{1*}, Ziguang Zhang¹, Anne Schreiber¹, Mark. A. Williams², Edward P. Morris¹ & David Barford¹

The ubiquitylation of cell-cycle regulatory proteins by the large multimeric anaphase-promoting complex (APC/C) controls sister chromatid segregation and the exit from mitosis^{1,2}. Selection of APC/C targets is achieved through recognition of destruction motifs, predominantly the destruction (D)-box³ and KEN (Lys-Glu-Asn)-box⁴. Although this process is known to involve a co-activator protein (either Cdc20 or Cdh1) together with core APC/C subunits^{1,2}, the structural basis for substrate recognition and ubiquitylation is not understood. Here we investigate budding yeast APC/C using single-particle electron microscopy and determine a cryo-electron microscopy map of APC/C in complex with the Cdh1 co-activator protein (APC/C^{Cdh1}) bound to a D-box peptide at ~10 Å resolution. We find that a combined catalytic and substrate-recognition module is located within the central cavity of the APC/C assembled from Cdh1, Apc10—a core APC/C subunit previously implicated in substrate recognition^{5–7}—and the cullin domain of Apc2. Cdh1 and Apc10, identified from difference maps, create a co-receptor for the D-box following repositioning of Cdh1 towards Apc10. Using NMR spectroscopy we demonstrate specific D-box–Apc10 interactions, consistent with a role for Apc10 in directly contributing towards D-box recognition by the APC/C^{Cdh1} complex. Our results rationalize the contribution of both co-activator and core APC/C subunits to D-box recognition^{8,9} and provide a structural framework for understanding mechanisms of substrate recognition and catalysis by the APC/C.

The APC/C is a multimeric E3 ubiquitin ligase assembled from 13 individual subunits^{1,2}. Many of the core proteins of the APC/C are comprised of multiple repeat motifs whose principle function is to provide a molecular scaffold, but whose exact biological role is not well understood. The best-characterized APC/C subunits are the cullin and RING proteins Apc2 and Apc11, which are responsible for catalytic activity, and the tetratricopeptide repeat (TPR) subunit Apc3/Cdc27, which interacts with a co-activator (either Cdc20 or Cdh1)^{10–12} and the APC/C subunit Apc10 (also known as Doc1)¹³. Both co-activator^{9,11,14–17} and core APC/C subunits^{5–9} have been implicated in substrate recognition, but the structural basis for this process is unknown. To address this question, we used single-particle electron microscopy (EM) to determine structures of budding yeast APC/C^{Cdh1} and substrates. The resultant EM maps are of excellent quality and detail. The maps show the characteristic triangular shape of the APC/C^{18–21} (Supplementary Fig. 1), but at higher resolution we visualize a lattice-like scaffold assembled from individual APC/C subunits defining a central cavity.

The APC/C co-activator Cdh1 was identified in negative-stain EM reconstructions as a prominent and discrete density feature present within the central cavity of APC/C^{Cdh1} and absent from APC/C (Fig. 1a, b). Its disc-shaped density, characteristic of an exposed WD40 β-propeller domain, is connected to the APC/C via an edge-on interface. Overall, with the exception of the Cdh1 density, APC/C and APC/C^{Cdh1} are similar, and the large conformational changes that

accompany co-activator binding to vertebrate APC/C^{19,21} are not evident. An ellipsoid-shaped density feature, resembling the β sandwich of Apc10 (refs 13, 22), situated adjacent to but not in contact with Cdh1, is more prominent in the presence of Cdh1 (Fig. 1a, b). Its close proximity to Cdh1 was intriguing in view of the role of Apc10 in contributing towards substrate recognition⁶, and the D-box-dependent processivity of the ubiquitylation reaction^{5,7}. To unequivocally identify Apc10 we generated APC/C^{ΔApc10} in complex with Cdh1 (APC/C^{ΔApc10–Cdh1}). The resultant APC/C^{ΔApc10–Cdh1} map showed complete loss of this ellipsoid density (Fig. 1c), confirming its identity as Apc10. Deletion of Apc10 also resulted in a depletion of Cdh1 density around the circumference of the β-propeller most distant from its contact to APC/C (Fig. 1c). Because deletion of Apc10 does not affect the APC/C subunit composition⁶ or abrogate Cdh1 binding (Supplementary Fig. 2), the partial loss of Cdh1 density is indicative of an increased flexibility of the WD40 domain of Cdh1. This finding and the reduced density of Apc10 in APC/C imply conformational interdependence of Apc10 and Cdh1.

To identify substrate-binding sites on APC/C^{Cdh1}, we used a fragment of Hsl1, a D-box (RxxLxxI/VxN)³ and KEN-box⁴-containing substrate with high affinity for APC/C^{Cdh1} (refs 14, 23). The ternary APC/C^{Cdh1–Hsl1} complex was catalytically competent, as judged by its ability to ubiquitylate Hsl1 (Supplementary Fig. 3a). Engagement of Hsl1 with APC/C^{Cdh1} is accompanied by a pronounced structural change involving Cdh1 and Apc10 (Fig. 1d). Specifically, the β-propeller domain of Cdh1 is bulkier, shifts ~7 Å towards Apc10, and new, well-defined density bridges Cdh1 to Apc10. Thus, Hsl1 promotes the formation of new connections between Cdh1 and Apc10, a result consistent with direct co-activator–substrate interactions^{9,11,14–17} and a role for Apc10 in mediating optimal substrate binding^{5–7,23}.

To define the specific roles of the D- and KEN-boxes in contributing to these conformational changes, we determined structures of APC/C^{Cdh1} in complex with synthetic peptides containing either a D-box or a KEN-box. Similar to previous results with D-box peptides^{11,24}, an 18-residue D-box peptide modelled on cyclin B (*Schizosaccharomyces pombe* Cdc13) completely inhibited APC/C^{Cdh1} activity towards Clb2 (a mitotic cyclin with D- and KEN-boxes) at 0.1 mM (Supplementary Fig. 4a). Figure 1e shows that D-box peptide generated similar structural changes to Hsl1; specifically, the WD40 domain of Cdh1 is shifted and new density connects it with Apc10 (Supplementary Movie 1). However, in contrast to the APC/C^{Cdh1–Hsl1} map, the extent of new density associated with Cdh1 is markedly reduced, indicating that the additional density in APC/C^{Cdh1–Hsl1} represents the larger Hsl1 substrate. Control experiments show that a mutant D-box peptide, which fails to bind APC/C^{Cdh1} (Supplementary Fig. 4c), induces no conformational changes (Supplementary Fig. 5). Binding of the KEN-box peptide to APC/C^{Cdh1} also promotes a repositioning of Cdh1 towards Apc10, but notably without the connecting density (Fig. 1f). This indicates that only D-box substrates promote a physical interconnection between Cdh1 and Apc10.

¹Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK. ²Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.

*These authors contributed equally to this work.

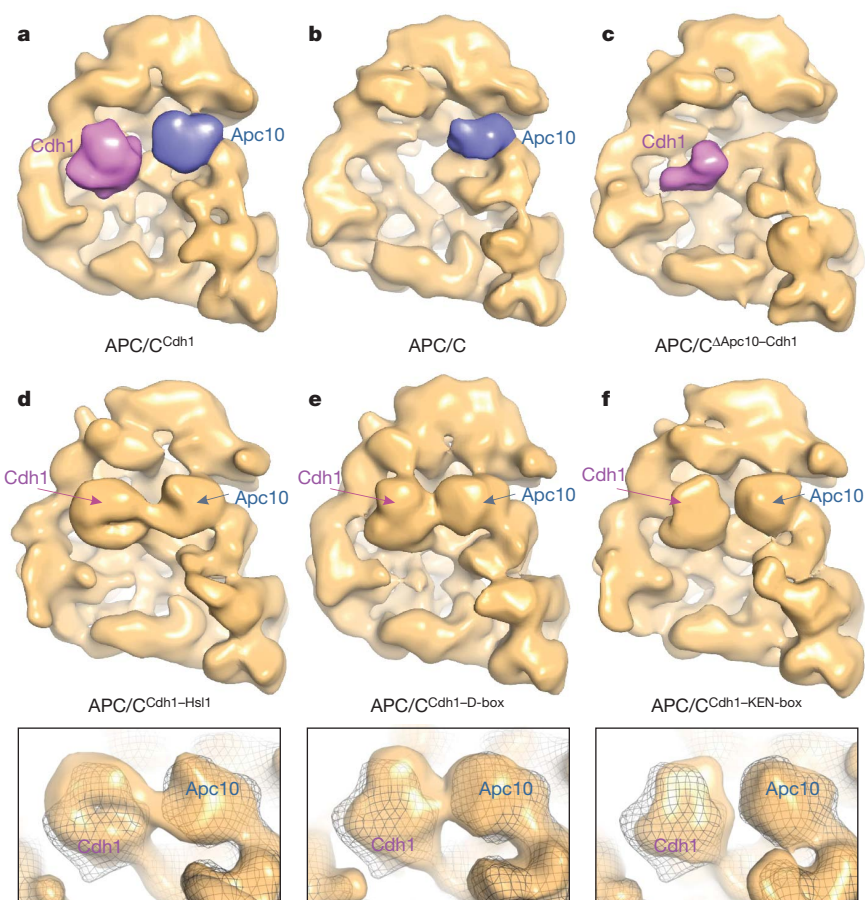


Figure 1 | Negative-stain EM reconstructions of budding yeast APC/C show that substrate binding to APC/C^{Cdh1} involves Cdh1 and Apc10. **a–c**, Molecular envelopes of APC/C^{Cdh1} (**a**), APC/C (**b**) and APC/C^{ΔApc10-Cdh1} (**c**). Density assigned to Cdh1 and Apc10 is shown in magenta and blue, respectively. The resolution of the APC/C^{Cdh1} binary complex is ~18–20 Å (Supplementary Fig. 10d). **d–f**, Negative-stain EM reconstructions of APC/C

C^{Cdh1}-Hsl1 complex (**d**), APC/C^{Cdh1}-D-box (**e**) and APC/C^{Cdh1}-KEN-box (**f**). Lower panels in **d**, **e** and **f** show details of the structural changes associated with Cdh1 and Apc10 in the presence of substrate compared with the superimposed binary APC/C^{Cdh1} map represented in mesh. Hsl1 and D-box and KEN-box peptides were used at saturating concentrations to promote stoichiometric APC/C^{Cdh1}-substrate ternary complexes.

To explore the structure of APC/C^{Cdh1-D-box} in more detail, we collected cryo-EM images of the complex and determined its structure at ~10 Å resolution. The cryo-EM map reproduces the overall features of the APC/C^{Cdh1-D-box} map generated from negatively stained particles, but with greatly enhanced detail and resolution (Fig. 2 and Supplementary Figs 6 and 7). Similar to the APC/C^{Cdh1-D-box} ternary complex obtained from negative-stain EM, the cryo-EM reconstruction shows density connecting Cdh1 and Apc10 (Figs 2 and 3). Docking the crystal structure of Apc10 (refs 13, 22) and the modelled

Cdh1 WD40 domain into their respective densities indicates further unassigned density linking Cdh1 to Apc10 (Fig. 3a, c). Notably, the best fit of Apc10 into the cryo-EM map positions a highly conserved loop required for D-box recognition⁷ adjacent to the density linking Apc10 with Cdh1. In contrast, residues on the opposite surface of Apc10 that contribute to APC/C interactions⁷ are oriented towards Apc2 (Fig. 3c).

These structural data revealing that Cdh1 and Apc10 become interconnected by bridging density in the presence of D-box substrates

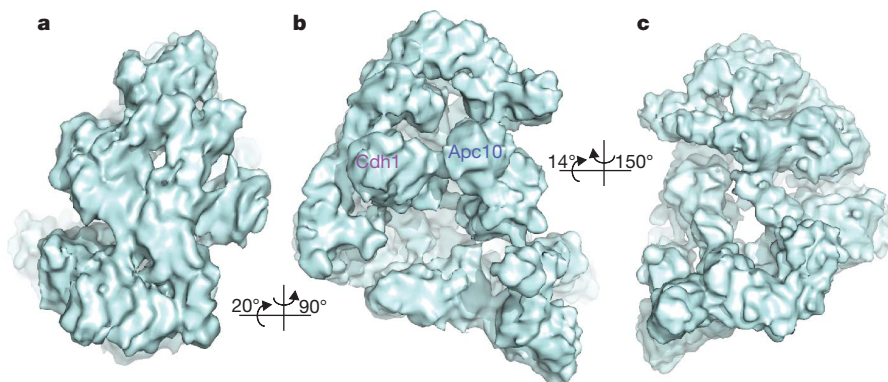


Figure 2 | Cryo-EM reconstruction of budding yeast APC/C^{Cdh1-D-box} reveals the lattice-like architecture of the complex. **a–c**, Three views of the complex with **b** similar to views shown in Fig. 1. Resolution is ~10 Å (Supplementary Fig. 12c).

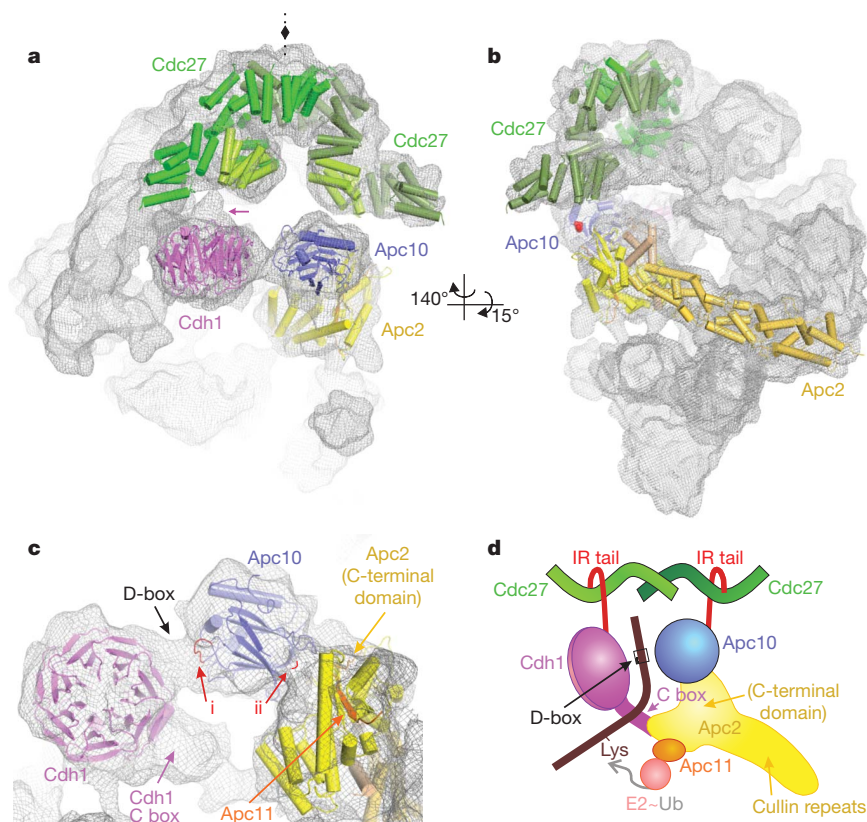


Figure 3 | Cdh1, Apc10, Apc2 and Apc11 form a substrate-recognition catalytic module. **a, b,** Two views of the cryo-EM APC/C^{Cdh1-D-box} complex. Protein density is represented by a mesh with fitted atomic coordinates of the Cdh1 β -propeller (modelled), Apc10 (ref. 22), Apc2–Apc11 (modelled on Cul4a–Rbx1 of SCF) and Cdc27 (ref. 26). Only the N-terminal β strand of Apc11 bound to the Apc2 C-terminal domain is modelled (orange). The two subunits of Cdc27 are shown in light and dark green. The view in **a** shows the two-fold symmetry axis of Cdc27. Density connecting Cdh1 to a TPR superhelix of the Cdc27 dimer is indicated by an arrow. TPR motifs 8 to 10 of Cdc27, implicated in IR-tail recognition²³, are shown in lighter colours. In **b**, the final residue of Apc10 observed in the crystal structure (Ser 256), 25 residues

rationalizes biochemical studies demonstrating that both co-activator and core APC/C subunits^{8,9,11,14–17}, specifically Apc10 (refs 5–7, 23), contribute to D-box-dependent recognition and processive ubiquitylation. The unassigned density bridging Apc10 and Cdh1 in the APC/C^{Cdh1-D-box} complex can be modelled as a D-box peptide, indicating that the binding site for the D-box is shared between the WD40 domain of Cdh1 and the β sandwich of Apc10. Cdh1 and Apc10 therefore generate a D-box co-receptor (Supplementary Fig. 8). Although biochemical data show that the D-box interacts with the conserved surface of the WD40 domain of the co-activator^{11,15}, direct interactions between D-box and Apc10 alone have not been previously demonstrated (unpublished data and ref. 7), possibly owing to the weak affinity of isolated Apc10 for D-box.

We used ¹H–¹⁵N-heteronuclear single quantum coherence (HSQC) NMR, a technique suitable for detecting weak protein–ligand interactions, to investigate potential Apc10–D-box interactions. The ¹H–¹⁵N-HSQC NMR spectrum of *Saccharomyces cerevisiae* Apc10, shown in Fig. 4, has a substantial number of well-dispersed peaks consistent with the Apc10 β -sandwich architecture²². However, the number of visible peaks is approximately half that expected for a 221-residue protein, and the visible peaks have a wide range of intensities. Reduced peak number and intensity variation are characteristic of proteins undergoing exchange between different conformational or oligomeric states. Spectra recorded with a twofold difference in protein concentration showed no change in position or shape of any dispersed peak, indicating that there is no sensitivity to any possible oligomerization

N-terminal to the IR motif, is indicated by red spheres. **c,** Details of the Cdh1 and Apc10 co-receptor for D-box. Both Cdh1 and Apc10 connect to Apc2. The N terminus of Cdh1, including the C box linking the WD40 domain to Apc2, is not modelled. The first red arrow (i) denotes the conserved loop (residues His 239 to Asp 244) of Apc10 implicated in D-box recognition⁷, and the second red arrow (ii) denotes the Lys 162 and Arg 163 of Apc10 responsible for APC/C affinity⁷. Two models for a possible fit of D-box to the density interconnecting Cdh1 and Apc10 are shown in Supplementary Fig. 8. **d,** Schematic of combined catalytic and substrate-recognition module responsible for D-box binding and substrate ubiquitylation. D-box is represented as binding to an interface between Cdh1 and Apc10.

equilibrium. Consequently, the features of the ¹H–¹⁵N-HSQC spectrum are best explained as a result of Apc10 adopting multiple conformations in intermediate to slow exchange (submillisecond to second timescales) in solution. Addition of a stoichiometric excess (~40-fold) of the D-box peptide used to generate the APC/C^{Cdh1-D-box} ternary complex resulted in more than 20 changes in amide peak position or relative intensity (Fig. 4). NMR-based measurement of the translational diffusion coefficient showed that the NMR-observed species is an Apc10 monomer of ~26 kDa. Thus, the changes in specific peaks on addition of peptide demonstrate that the D-box peptide interacts with monomeric Apc10, altering the chemical environment and/or the conformational equilibrium of a subset of its residues. However, the low intensity and proportion of visible amide peaks made sequential assignment and full characterization of the D-box binding site on Apc10 impracticable.

To establish whether the peptide-induced changes of the Apc10 NMR spectrum are specifically D-box dependent, we performed a series of control experiments. First, a different D-box peptide (a 19-residue peptide modelled on *S. cerevisiae* Clb2 whose sequence identity with Cdc13 is confined to the D-box) produced very similar NMR spectral changes to the Cdc13 D-box (Fig. 4). Second, a mutant D-box Cdc13 peptide resulted in only minor changes in the Apc10 NMR spectrum, consistent with greatly reduced binding. Finally, the Hsl1 KEN-box peptide that, from the APC/C^{Cdh1-KEN-box} EM analysis, does not bridge Cdh1 and Apc10, resulted in an essentially identical spectrum to that of the apoprotein, with none of the changes seen for

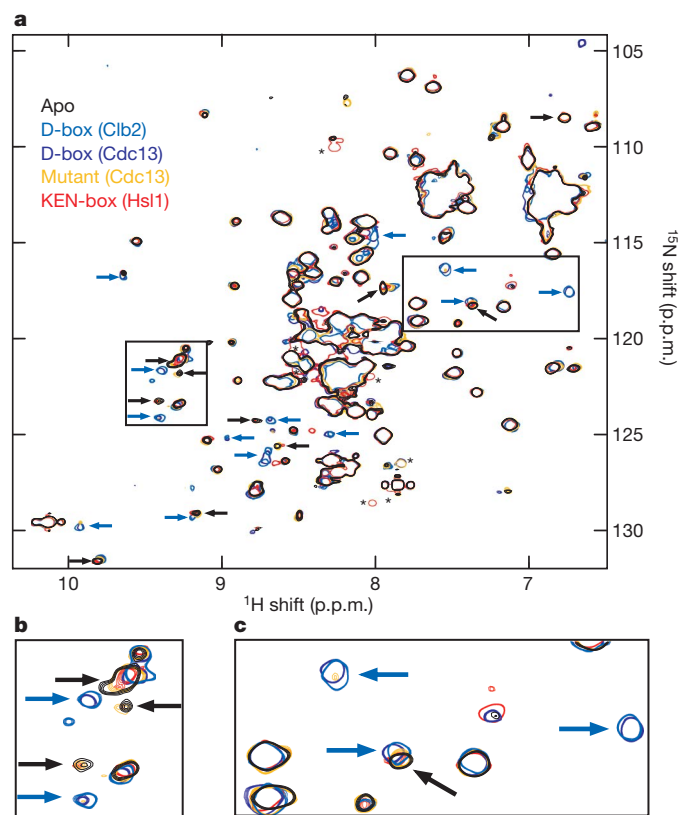


Figure 4 | ^1H - ^{15}N HSQC spectra of Apc10. **a–c**, Overlaid are spectra of the apoprotein and protein in the presence of stoichiometric excess of each of four peptides. The complete amide region (**a**) and, for clarity, expanded views of two boxed sub-regions (**b**, **c**) are shown. Spectra in the presence of either of the two D-box-containing peptides show common changes with respect to the apoprotein spectrum, namely absence of the peaks seen in the apoprotein (black arrows) and new or shifted peaks not seen in the apospectrum (blue arrows). In contrast, spectra in the presence of either the Cdc13-derived peptide in which four residues of the D-box motif are mutated to alanine or a peptide containing a KEN-box motif are very similar to the apoprotein, retaining all of the peaks marked by black arrows. The spectrum with the mutant Cdc13 peptide does in some cases show low-intensity peaks at the positions indicated by blue arrows (see expanded views in **b** and **c**), indicating a very weak residual interaction. These spectra are consistent with a D-box-dependent interaction with Apc10. Peaks arising from natural abundance ^{15}N amides in the unbound peptide that are protected from solvent exchange are indicated by an asterisk.

the two D-box-containing peptides. These NMR data therefore provide strong evidence for a direct interaction between Apc10 and D box, supporting the notion that Apc10 participates in D-box recognition.

To gain further insight into the mechanisms of substrate recognition and ubiquitylation, we modelled atomic structures of Apc2 and Cdc27 into the molecular envelope of the APC/C^{Cdh1-D-box} map. We fitted a homology model of Apc2, based on Cul4a-Rbx1, allowing for small adjustments of the carboxy-terminal domain relative to the cullin repeats (Fig. 3 and Supplementary Figs 7 and 9). Continuous density attaches the globular C-terminal domain to that of the cullin repeats, which are seen as a long stalk-like density that transverses one side of the complex (Fig. 3 and Supplementary Fig. 7). The APC/C^{Cdh1-D-box} cryo-map reveals that Cdh1 and Apc10 are both connected to the Apc2 C-terminal domain (Fig. 3c and Supplementary Fig. 7). Notably, the interaction of the C-terminal domain of Apc2 with substrate adaptor subunits contrasts with the Skp1-cullin-F-box (SCF) complex in which the amino-terminal cullin repeat of Cul1 interacts with substrate adaptors²⁵.

Cdc27 is a dimer and we docked its N-terminal dimerization domain²⁶ into the globular structure at the head of the TPR sub-complex, and independently positioned the modelled C-terminal

TPR superhelices of the Cdc27 subunits into the curved tubular densities extending from the globular domain (Fig. 3a, b and Supplementary Fig. 7), consistent with the mapping of Cdc27 (unpublished data). Although not imposed in the fitting, these docked TPR superhelices are related by the same dyad symmetry as the Cdc27 dimerization domain, therefore preserving the overall two-fold symmetry of Cdc27 (Fig. 3a). The organization of Apc2 and Cdc27 in close proximity to Cdh1 and Apc10 visualized in our APC/C^{Cdh1-D-box} structure unifies previous models of APC/C subunit topologies^{10,12,13,23} (Fig. 3d). Cdh1 is known to interact through its C-terminal Ile-Arg (IR) tail with Cdc27^{10–12}, and in *S. cerevisiae*, Cdh1 also requires Apc2 for optimal binding¹². The structures fitted to the EM map show that with the C terminus of Cdh1 in contact with Cdc27, its N-terminal C box is positioned to contact Apc2 (Fig. 3 and Supplementary Fig. 7)^{12,17}. Pull-down experiments on recombinant human proteins have shown that Apc10 interacts with Cdc27 through its C-terminal region, which also contains an IR motif¹³, whereas in *S. cerevisiae*, Apc10 associates preferentially with a sub-complex of Apc1, Apc2 and Apc11 (ref. 12). Our EM data position Apc10 close to the second Cdc27 subunit. Consequently, the human and yeast biochemical data are explained by the extensive interface between Apc10 and Apc2, and the flexible C-terminal IR tail of Apc10 binding to the Cdc27 TPR superhelix.

This study identifies Cdh1 and Apc10 as a co-receptor for D-box. Individually, co-activator and APC/C possess low affinity and specificity for substrate⁹ and therefore cooperatively enhance substrate affinity through multivalency. Definition of the subunit organization and generation of a pseudo-atomic structure of the APC/C (unpublished data), together with characterization of the D-box co-receptor presented here, provide the conceptual framework for a mechanistic understanding of the APC/C.

METHODS SUMMARY

Generation of APC/C and complexes with Cdh1 and substrates. APC/C and APC/C^{Apc10} were isolated from *S. cerevisiae* and ubiquitylation assays were performed essentially as described⁶. *S. cerevisiae* His₆-Cdh1 was expressed in *Spodoptera frugiperda* (Sf9) cells and purified using nickel-nitrilotriacetic acid (Ni-NTA). APC/C^{Cdh1} was prepared by loading excess Cdh1 onto APC/C immobilized on calmodulin resin, and eluted as for APC/C. APC/C^{Cdh1}-substrate complexes were generated as described in Methods.

EM and image analysis. Purified APC/C (~0.2 mg ml⁻¹) from peak elution fractions was applied to Quantifoil 1.2- or 2-μm aperture grids coated with continuous thin carbon and either negatively stained for EM at room temperature (20 °C) or flash frozen using a Vitrobot for cryo-EM. Images were recorded in an FEI TF20 electron microscope under low-dose conditions using a Tietz F415 CCD camera. Three-dimensional maps were calculated from molecular images using programs from Imagic²⁷, Spider²⁸ and EMAN²⁹.

NMR analysis. ^1H - ^{15}N HSQC spectra were recorded at 25 °C over 5.5 or 11 h for samples of Apc10 alone and in the presence of the four peptide samples using a (^1H , ^{15}N , ^{13}C) triple resonance cryoprobe on a 700 MHz Bruker Avance III spectrometer. Spectra were processed identically and displayed to compensate for concentration and/or recording time differences.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 November 2009; accepted 1 November 2010.

Published online 24 November 2010.

- Peters, J. M. The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nature Rev. Mol. Cell Biol.* **7**, 644–656 (2006).
- Thornton, B. R. & Toczyski, D. P. Precise destruction: an emerging picture of the APC. *Genes Dev.* **20**, 3069–3078 (2006).
- Glotzer, M., Murray, A. W. & Kirschner, M. W. Cyclin is degraded by the ubiquitin pathway. *Nature* **349**, 132–138 (1991).
- Pfleger, C. M. & Kirschner, M. W. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.* **14**, 655–665 (2000).
- Carroll, C. W. & Morgan, D. O. The Doc1 subunit is a processivity factor for the anaphase-promoting complex. *Nature Cell Biol.* **4**, 880–887 (2002).
- Passmore, L. A. *et al.* Doc1 mediates the activity of the anaphase-promoting complex by contributing to substrate recognition. *EMBO J.* **22**, 786–796 (2003).
- Carroll, C. W., Enquist-Newman, M. & Morgan, D. O. The APC subunit Doc1 promotes recognition of the substrate destruction box. *Curr. Biol.* **15**, 11–18 (2005).

8. Yamano, H., Gannon, J., Mahbubani, H. & Hunt, T. Cell cycle-regulated recognition of the destruction box of cyclin B by the APC/C in *Xenopus* egg extracts. *Mol. Cell* **13**, 137–147 (2004).
9. Eytan, E., Moshe, Y., Braunstein, I. & Herskho, A. Roles of the anaphase-promoting complex/cyclosome and of its activator Cdc20 in functional substrate binding. *Proc. Natl Acad. Sci. USA* **103**, 2081–2086 (2006).
10. Vodermaier, H. C., Gieffers, C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. TPR subunits of the anaphase-promoting complex mediate binding to the activator protein CDH1. *Curr. Biol.* **13**, 1459–1468 (2003).
11. Kraft, C., Vodermaier, H. C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. The WD40 propeller domain of Cdh1 functions as a destruction box receptor for APC/C substrates. *Mol. Cell* **18**, 543–553 (2005).
12. Thornton, B. R. *et al.* An architectural map of the anaphase-promoting complex. *Genes Dev.* **20**, 449–460 (2006).
13. Wendt, K. S. *et al.* Crystal structure of the APC10/DOC1 subunit of the human anaphase-promoting complex. *Nature Struct. Biol.* **8**, 784–788 (2001).
14. Burton, J. L. & Solomon, M. J. D box and KEN box motifs in budding yeast Hsl1p are required for APC-mediated degradation and direct binding to Cdc20p and Cdh1p. *Genes Dev.* **15**, 2381–2395 (2001).
15. Hilioti, Z., Chung, Y. S., Mochizuki, Y., Hardy, C. F. & Cohen-Fix, O. The anaphase inhibitor Pds1 binds to the APC/C-associated protein Cdc20 in a destruction box-dependent manner. *Curr. Biol.* **11**, 1347–1352 (2001).
16. Pfeleger, C. M., Lee, E. & Kirschner, M. W. Substrate recognition by the Cdc20 and Cdh1 components of the anaphase-promoting complex. *Genes Dev.* **15**, 2396–2407 (2001).
17. Schwab, M., Neutzner, M., Mockler, D. & Seufert, W. Yeast Hct1 recognizes the mitotic cyclin Clb2 and other substrates of the ubiquitin ligase APC. *EMBO J.* **20**, 5165–5175 (2001).
18. Passmore, L. A. *et al.* Structural analysis of the anaphase-promoting complex reveals multiple active sites and insights into polyubiquitylation. *Mol. Cell* **20**, 855–866 (2005).
19. Dube, P. *et al.* Localization of the coactivator Cdh1 and the cullin subunit Apc2 in a cryo-electron microscopy model of vertebrate APC/C. *Mol. Cell* **20**, 867–879 (2005).
20. Ohi, M. D. *et al.* Structural organization of the anaphase-promoting complex bound to the mitotic activator Slp1. *Mol. Cell* **28**, 871–885 (2007).
21. Herzog, F. *et al.* Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science* **323**, 1477–1481 (2009).
22. Au, S. W., Leng, X., Harper, J. W. & Barford, D. Implications for the ubiquitination reaction of the anaphase-promoting complex from the crystal structure of the Doc1/Apc10 subunit. *J. Mol. Biol.* **316**, 955–968 (2002).
23. Matyskiela, M. E. & Morgan, D. O. Analysis of activator-binding sites on the APC/C supports a cooperative substrate-binding mechanism. *Mol. Cell* **34**, 68–80 (2009).
24. Yamano, H., Tsurumi, C., Gannon, J. & Hunt, T. The role of the destruction box and its neighbouring lysine residues in cyclin B for anaphase ubiquitin-dependent proteolysis in fission yeast: defining the D-box receptor. *EMBO J.* **17**, 5670–5678 (1998).
25. Zheng, N. *et al.* Structure of the Cul1–Rbx1–Skp1–F box^{Skp2} SCF ubiquitin ligase complex. *Nature* **416**, 703–709 (2002).
26. Zhang, Z. *et al.* Molecular structure of the N-terminal domain of the APC/C subunit Cdc27 reveals a homo-dimeric tetratricopeptide repeat architecture. *J. Mol. Biol.* **397**, 1316–1328 (2010).
27. van Heel, M. *et al.* Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.* **33**, 307–369 (2000).
28. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
29. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by a Cancer Research UK grant to D.B. We thank F. Beuron for help with the early stages of this project and for EM support. We are grateful to J. Kirkpatrick for the use of the facilities of the UCL/Birkbeck Institute of Structural Molecular Biology (ISMB) Biomolecular NMR Centre.

Author Contributions All authors contributed to experimental design, data analysis and manuscript preparation. P.C.A.d.F. and E.H.K. collected and analysed EM data, E.H.K. prepared APC/C samples and performed ubiquitylation assays. P.C.A.d.F. determined the three-dimensional EM reconstructions and fitted coordinates. M.A.W. performed NMR experiments and analysed NMR data. E.P.M. helped collect and analyse EM data.

Author Information EM maps have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMD-1815 (cryo-EM APC/C^{Cdh1–D-box}), EMD-1816 (apo APC/C), EMD-1817 (APC/C^{Cdh1}), EMD-1818 (APC/C^{Cdh1–KEN-box}) and EMD-1819 (APC/C^{Cdh1–Hsl1}). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.B. (david.barford@icr.ac.uk).

METHODS

Generation of APC/C and complexes with Cdh1 and substrates. APC/C and APC/C^{ΔApc10} were isolated from *S. cerevisiae* and ubiquitylation assays were performed as described^{6,30} except that the calmodulin resin elution buffer was 25 mM HEPES (pH 8.0), 150 mM NaCl, 1 mM MgCl₂, 2 mM EGTA, 3 mM tris(2-carboxyethyl)phosphine (TCEP), and 0.03% (v/v) n-dodecyl-β-D-maltoside (DDM). Peak elution fractions were used for EM analysis. *S. cerevisiae* His₆-Cdh1 was expressed in Sf9 cells and purified using Ni-NTA. APC/C^{Cdh1} was prepared by loading excess Cdh1 onto APC/C immobilized on calmodulin resin, thereby ensuring formation of a stoichiometric APC/C^{Cdh1} complex, and eluted as for APC/C. Association of Cdh1 to APC/C was confirmed by SDS-PAGE and western blotting analyses (Supplementary Fig. 2), and by E3 ligase assays showing that APC/C^{Cdh1} ubiquitylated Hsl1 and Clb2 (Supplementary Figs 3 and 4). Hsl1^{667–872} was expressed in BL21(DE3) RIL cells and purified by Ni-NTA and gel-filtration chromatography and added to APC/C^{Cdh1} to a final concentration of 1.5 μM for EM data collection, greatly in excess of the APC/C^{Cdh1} concentration (Supplementary Fig. 3a). Previous work had shown that Hsl1 forms a stable 1:1 complex with APC/C^{Cdh1} at ~0.05 μM (ref. 23). APC/C^{Cdh1-Hsl1} was completely inhibited towards Clb2 (Supplementary Fig. 3b). D-box peptide inhibited APC/C^{Cdh1} ubiquitylation of Clb2 at 0.1 mM (Supplementary Fig. 4a), similar to previous findings^{11,24}. APC/C^{Cdh1-D-box} and APC/C^{Cdh1-AD-box} were prepared by adding peptide to APC/C^{Cdh1} to a final concentration of 0.3 mM. KEN-box peptide inhibited APC/C^{Cdh1} ubiquitylation of Clb2 at 1 mM (Supplementary Fig. 4b) and was therefore used at 10 mM for the APC/C^{Cdh1-KEN-box} structure. Peptides used in the EM structural analysis and ubiquitylation assays were as follows. D-box, NVPKKRHALDDVSNFHNK; AD-box, NVPKKAHAADDVSAFHNK; KEN-box, GVSTNKENEGPEYPTKIK KEHQK (D-box, mutant D-box and KEN-box underlined). D-box and KEN-box peptides were modelled on *S. pombe* Cdc13 and *S. cerevisiae* Hsl1, respectively. Stock solutions were dissolved at 10–20 mM in 100 mM Tris HCl (pH 8.0). For competitions assays, peptides were used at the final stated concentrations.

EM of negative-stained samples. Purified APC/C and its Cdh1 and substrate complexes at ~0.2 mg ml⁻¹ were applied to Quantifoil 2/2 EM grids coated with a second layer of thin carbon. The grids were negatively stained with 2% (w/v) uranyl acetate. The samples were imaged at room temperature (20 °C) in an FEI Tecnai TF20 electron microscope at an accelerating voltage of 200 kV, in low-dose mode with an exposure of ~100 e⁻ Å⁻², a nominal magnification of ×50,000 and an underfocus of ~1.2 μm, giving rise to a first minimum in the contrast transfer function at ~17 Å. Images were recorded using a Tietz F415 (4k × 4k) CCD camera and adjacent boxes of 2 × 2 pixels were averaged, resulting in a calibrated sampling of 3.47 Å pixel⁻¹. The images recorded for all negatively stained samples were consistent with that of APC/C^{Cdh1} shown in Supplementary Fig. 10a, including those of samples of APC/C^{ΔApc10-Cdh1} (Supplementary Fig. 11).

Cryo-EM. Samples of purified APC/C^{Cdh1-D-box} were applied to Quantifoil 1.2/1.3 EM grids coated with a second layer of thin carbon, blotted and plunged into liquid ethane using an FEI Vitrobot. The grids were transferred into a Tecnai TF20 and maintained at approximately -178 °C using a Gatan 626 cryo-holder. Images were recorded in a similar way to that described for negatively stained samples, except that focal pairs were recorded at an underfocus of ~2.5 μm and ~4 μm, using an electron dose of ~20 e⁻ Å⁻² (for each exposure) and a nominal magnification of ×63,000, resulting in a sampling of 2.82 Å pixel⁻¹. The first recorded CCD images of each focal pair (closer to focus) were carefully screened and only those with a power spectrum showing Thon rings extending isotropically beyond 10 Å were selected for further analysis.

Image analysis of negatively stained samples. Image processing was performed using Imagic²⁷, Spider²⁸ and EMAN²⁹ programs. Image processing was initiated with the analysis of the APC/C^{Cdh1} complex. Molecular images were manually selected (Supplementary Fig. 10a) using the EMAN boxer software in order to assemble a data set ultimately formed of 12,529 images. A preliminary evaluation of the resulting data set was carried out by calculating reference-free image-class averages using the refine2d routine from EMAN. Three classes, which were judged to be approximately mutually orthogonal, were selected from the preliminary set for angular assignment using the Imagic C1 start-up procedure. These were used to assign angles, by angular reconstitution, to a further selection of 112 classes, which were subsequently back-projected in order to create an *ab initio* three-dimensional map. This map was used as the first reference for refinement using a combination of Imagic and Spider software. The refinement consisted of multiple rounds of multi-reference alignment, classification, angular assignment (to selected image-class averages) by projection matching and three-dimensional reconstruction by back-projection. In the last round of refinement a total of 4,000 class averages were calculated, of which 1,433 were selected to calculate the final three-dimensional map. Examples of class averages used in the reconstruction and their respective reprojections are shown in Supplementary Fig. 10b. The angular distribution of the classes used in the final reconstruction is shown in

Supplementary Fig. 10c. The resolution of the final map of APC/C^{Cdh1} was estimated by Fourier shell correlation as 18–20 Å, depending on the resolution criteria (Supplementary Fig. 10d). Negatively stained APC/C^{ΔApc10-Cdh1} appear indistinguishable from APC/C^{Cdh1} (Supplementary Fig. 11).

The final map calculated for APC/C^{Cdh1} was used as a starting reference for the analysis of all other negatively stained APC/C complexes, followed by the same refinement procedures. The total number of molecular images used in the image analysis of each sample is summarized in Supplementary Table 1. Representations of the maps were generated using PyMOL (<http://www.pymol.org>).

Image analysis of data from cryo-EM. The contrast transfer function (CTF) was measured for each CCD image selected for analysis and corrected by phase reversal. A data set of 9,474 molecular images of the APC/C^{Cdh1-D-box} complex was assembled manually using the EMAN boxer software, from the first recorded CCD image of each focal pair (closer to focus, Supplementary Fig. 12a), using the higher contrast second image as an aid for the selection. The subsequent analysis was performed using Imagic and Spider software. The map of the APC/C^{Cdh1-D-box} complex determined by the analysis of negatively stained samples was used as a starting reference for the analysis. The molecular images were aligned and their angular assignment performed by projection matching against the initial reference map. A first three-dimensional reconstruction was calculated by back-projection and this was further refined by multiple rounds of alignment, angular assignment by projection matching and back-projection. The angular distribution of the images for the calculation of the final map is shown in Supplementary Fig. 12b. The resolution of the final map, estimated by Fourier shell correlation, is 9–10 Å, depending on the resolution criteria (Supplementary Fig. 12c). For the representation of the final reconstruction a reverse B factor of -300 was applied, in order to optimize the agreement between the resulting reconstruction and the fitted coordinates, followed by a Fourier low-pass filtration to 9.5 Å. PyMOL (<http://www.pymol.org>) was used to generate the representations of the map.

Fitting atomic coordinates to cryo-EM map of APC/C^{Cdh1-D-box}. Apc10 is based on *S. cerevisiae* Apc10/Doc1 (PDB code 1GQP)²², the N-terminal dimerization domain of Cdc27 is based on *E. coli* Cdc27 (PDB code 3KAE)²⁶ and the C-terminal TPR superhelix is based on the model in ref. 26 (overall sequence identity of 16%). *S. cerevisiae* Cdh1 and Apc2 were modelled using the PHYRE server³¹ based on coordinates (PDB codes 2GNQ (WDR5)³² and 2HYE (Cul4a-Rbx1)³³, respectively, with overall sequence identities of 17% and 11%).

Atomic coordinates of Apc10 (PDB code 1GQP) and the N-terminal homodimerization domain of Cdc27 (PDB code 3KAE) and the molecular models of Cdh1, Apc2 (cullin domain and cullin repeats independently) and two copies of the model of the C-terminal TPR repeats of Cdc27 were docked into the cryo-EM map of the APC/C^{Cdh1-D-box} complex using URO software³⁴ (correlation coefficient of 0.82). The fitted coordinates were converted to densities, Fourier low-pass filtered to 9.5 Å and rendered to yield a volume corresponding to their calculated molecular mass of 243 kDa, assuming a protein density of 0.844 Da Å⁻³. The filtered coordinates were used to guide the rendering of the APC/C^{Cdh1-D-box} map, resulting in a volume corresponding to ~1.13 MDa. Furthermore, the comparison of the level of detail shown by the docked coordinates and that in our three-dimensional map of APC/C^{Cdh1-D-box} determined from cryo-EM data, is supportive of a resolution estimate of ~10 Å (Supplementary Fig. 7).

The protocol by which the *ab initio* APC/C^{Cdh1} map was calculated, which was the initial reference for the analysis of all complexes presented here, results in three-dimensional maps with ambiguity with respect to their hand. However, the hand of the APC/C complex as presented here has been previously determined by random conical tilt methods^{19,21}. In the present work the hand shown is supported by the agreement between the docked coordinates and their respective densities.

NMR analysis. Uniformly ¹⁵N-labelled Apc10 was purified from *E. coli* grown in a defined minimal medium supplemented with ¹⁵N-ammonium sulphate using constructs and protocols previously described²². Peptides (Supplementary Table 2) were dissolved in 100 mM Tris/MOPS to the lower limit of either their maximum solubility or a 100 mM concentration, and their pH was adjusted to ~8 with NaOH. Peptide was added to protein stock to a final concentration of 5 mM. Protein solubility and propensity to aggregation determined the optimal solution conditions for NMR data collection. All NMR samples were in 90% H₂O:10% D₂O, 77–85 mM NaCl, 4.5 mM DTT, 90 mM Tris/MOPS buffer pH 8.0. For the spectra shown, with the exception of the Clb2 sample, final protein concentration was 130–160 μM. Addition of the Clb2 D-box-containing peptide caused substantial precipitation of the protein (also seen to a lesser degree with the Cdc13 D-box peptide), leading to a final protein concentration in this sample of 64 μM. The pH of final protein-peptide mixtures was confirmed by NMR chemical shift of Tris methylene peaks³⁵ to be 8.0 ± 0.1. The Hsl1 KEN-box peptide used for NMR studies inhibited APC/C^{Cdh1}-catalysed ubiquitylation of Clb2 at a concentration of 2 mM (data not shown).

¹H-¹⁵N HSQC spectra of 1,024 × 128 complex points were recorded for each sample using a (¹H, ¹⁵N, ¹³C) triple resonance cryoprobe on a 700 MHz Bruker

Avance III spectrometer with identical spectral widths. Data were recorded at 25 °C for 5.5 or 11 h. The same spectral processing was applied to each spectrum (Gaussian apodization in ^1H and sine-bell apodization in ^{15}N dimensions and zero filling to $2,048 \times 512$ points before Fourier transformation and polynomial baseline correction) using NMRPipe³⁶. Spectra were overlaid in CCPNmr Analysis³⁷ and contour levels matched for concentration and recording time differences using the intense peaks common to all five spectra.

The field-gradient dependence of the signal intensity of the central region of the ^{15}N -edited spectrum (containing the strongest signals) of the APC10 with the Cdc13 D-box peptide was used to measure the extent of translational diffusion during a fixed time interval³⁸. The data fit a model corresponding to a single species of molecular mass ~ 26 kDa, that is, that of the monomer, with no indication of a significant NMR-observable population of dimer or higher-order oligomers.

30. Passmore, L. A., Barford, D. & Harper, J. W. Purification and assay of the budding yeast anaphase-promoting complex. *Methods Enzymol.* **398**, 195–219 (2005).
31. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols* **4**, 363–371 (2009).
32. Schuetz, A. *et al.* Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO J.* **25**, 4245–4252 (2006).
33. Angers, S. *et al.* Molecular architecture and assembly of the DDB1–CUL4A ubiquitin ligase machinery. *Nature* **443**, 590–593 (2006).
34. Navaza, J., Lepault, J., Rey, F. A., Alvarez-Rua, C. & Borge, J. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1820–1825 (2002).
35. Baryshnikova, O. K., Williams, T. C. & Sykes, B. D. Internal pH indicators for biomolecular NMR. *J. Biomol. NMR* **41**, 5–7 (2008).
36. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
37. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
38. Wu, D., Chen, A. & Johnson, C. S. Jr. An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *J. Magn. Reson. A* **115**, 260–264 (1995).

A unique chromatin signature uncovers early developmental enhancers in humans

Alvaro Rada-Iglesias¹, Ruchi Bajpai¹, Tomek Swigut¹, Samantha A. Brugmann¹, Ryan A. Flynn¹ & Joanna Wysocka^{1,2}

Cell-fate transitions involve the integration of genomic information encoded by regulatory elements, such as enhancers, with the cellular environment^{1,2}. However, identification of genomic sequences that control human embryonic development represents a formidable challenge³. Here we show that in human embryonic stem cells (hESCs), unique chromatin signatures identify two distinct classes of genomic elements, both of which are marked by the presence of chromatin regulators p300 and BRG1, monomethylation of histone H3 at lysine 4 (H3K4me1), and low nucleosomal density. In addition, elements of the first class are distinguished by the acetylation of histone H3 at lysine 27 (H3K27ac), overlap with previously characterized hESC enhancers, and are located proximally to genes expressed in hESCs and the epiblast. In contrast, elements of the second class, which we term 'poised enhancers', are distinguished by the absence of H3K27ac, enrichment of histone H3 lysine 27 trimethylation (H3K27me3), and are linked to genes inactive in hESCs and instead are involved in orchestrating early steps in embryogenesis, such as gastrulation, mesoderm formation and neurulation. Consistent with the poised identity, during differentiation of hESCs to neuroepithelium, a neuroectoderm-specific subset of poised enhancers acquires a chromatin signature associated with active enhancers. When assayed in zebrafish embryos, poised enhancers are able to direct cell-type and stage-specific expression characteristic of their proximal developmental gene, even in the absence of sequence conservation in the fish genome. Our data demonstrate that early developmental enhancers are epigenetically pre-marked in hESCs and indicate an unappreciated role of H3K27me3 at distal regulatory elements. Moreover, the wealth of new regulatory sequences identified here provides an invaluable resource for studies and isolation of transient, rare cell populations representing early stages of human embryogenesis.

Recent reports demonstrated that active enhancers can be identified by epigenomic profiling of p300 (ref. 4), H3K4me1 and H3K27ac^{5,6}. To characterize the enhancer repertoire of hESCs we performed chromatin immunoprecipitation coupled to massively parallel DNA sequencing (ChIP-seq) using antibodies recognizing chromatin regulators (that is, p300, BRG1) and histone modifications (that is, H3K4me1, H3K27ac, H3K4me3, H3K27me3) that distinguish distal elements from proximal promoters^{5,6} (Supplementary Fig. 1). As expected, previously characterized hESC enhancers (for example, *NANOG* (ref. 7) and *OCT4* (also called *POU5F1*)⁸) were bound by p300 and flanked by H3K4me1 and H3K27ac marked chromatin, but were not enriched for H3K27me3 or H3K4me3 (Fig. 1a and Supplementary Fig. 2a). Genome-wide analysis defined 5,118 genomic regions (hereafter referred to as class I elements) marked by a similar chromatin signature (that is, high p300, H3K4me1 and H3K27ac, low, if any, H3K4me3, and absence of H3K27me3), representing putative active hESC enhancers (Fig. 1b and Supplementary Data 1).

Interestingly, in the vicinity of many early developmental genes we noted promoter-distal p300-bound regions that were marked by

H3K4me1 but, in contrast to the active hESC enhancers, lacked H3K27ac and were instead enriched for H3K27me3, a modification associated with polycomb silencing⁹ (Fig. 1a). Overall, we identified 2,287 p300-bound regions devoid of H3K27ac and marked by H3K27me3, which we will hereafter refer to as class II elements (Fig. 1b and Supplementary Data 1). In general, class II elements showed enrichment of both H3K27me3 and H3K4me1 flanking p300 peaks (Fig. 1b). In contrast, analysis of previously described adult tissue-specific enhancers^{10–13} revealed no enrichment for any of the interrogated modifications (Supplementary Fig. 2b–e).

p300 enrichment levels were comparable at class I and II elements (Supplementary Fig. 3a), both classes were bound by BRG1 (Supplementary Fig. 3b), and showed similar genomic distribution relative to annotated transcription start sites (TSS), with over 95% of regions located away from promoters (Fig. 1c). Moreover, only 1.7% and 3.9% of class I and class II elements, respectively, overlapped with CpG islands, in sharp contrast to the 50% overlap observed for promoters. Another property of enhancers is their relative nucleosomal depletion compared to the flanking regions^{14,15}. Using FAIRE-seq (formaldehyde-assisted isolation of regulatory elements¹⁶ coupled to sequencing) we showed that class I and II elements were comparably nucleosome-depleted (Supplementary Fig. 3c). Furthermore, examination of a reported DNA-methylation-sensitive restriction enzyme data set from hESCs¹⁷ revealed similar levels of DNA hypomethylation at class I and class II elements (Supplementary Fig. 3d).

ChIP-seq results were validated by ChIP-qPCR at a representative subset of class I and class II elements (labelled with the name of their closest gene) (Supplementary Figs 4a–d and 5). Further examination of the H3K27ac and H3K27me3 enrichments showed a mutually exclusive marking pattern at class I and class II elements (Supplementary Fig. 6). Sequential ChIP-qPCR demonstrated a simultaneous presence of H3K4me1/K27ac at class I regions, and H3K4me1/K27me3 at class II regions, indicating that the concurrent enrichments of H3K4me1 and H3K27me3 were not due to cell population heterogeneity (Fig. 2a, b). Moreover, consistent with H3K27me3, we observed enrichment of the PRC2 component, SUZ12, at class II elements (Supplementary Fig. 4e). We also detected preferential association of RNA POL2 with class I elements, as compared to class II elements, including its unphosphorylated, Ser 5 phosphorylated and Ser 2 phosphorylated forms (Supplementary Fig. 7a–c).

Next we asked whether transcriptional status of nearby genes differs between the two classes. To this end, we analysed hESC transcriptome by RNA-seq and examined transcripts originating from TSS closest to the elements of each class. Class-I-associated gene expression was significantly higher than expression of all genes, or of class-II-associated genes, which were poorly expressed (Fig. 2c). In agreement, class-II-associated TSS were enriched for both H3K27me3 and H3K4me3, whereas class-I-associated TSS were marked by high H3K4me3 levels (Supplementary Fig. 8a, c). Thus, the two classes defined by unique chromatin signatures are also distinguished by the transcriptional status of associated genes.

¹Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California 94305, USA. ²Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA.

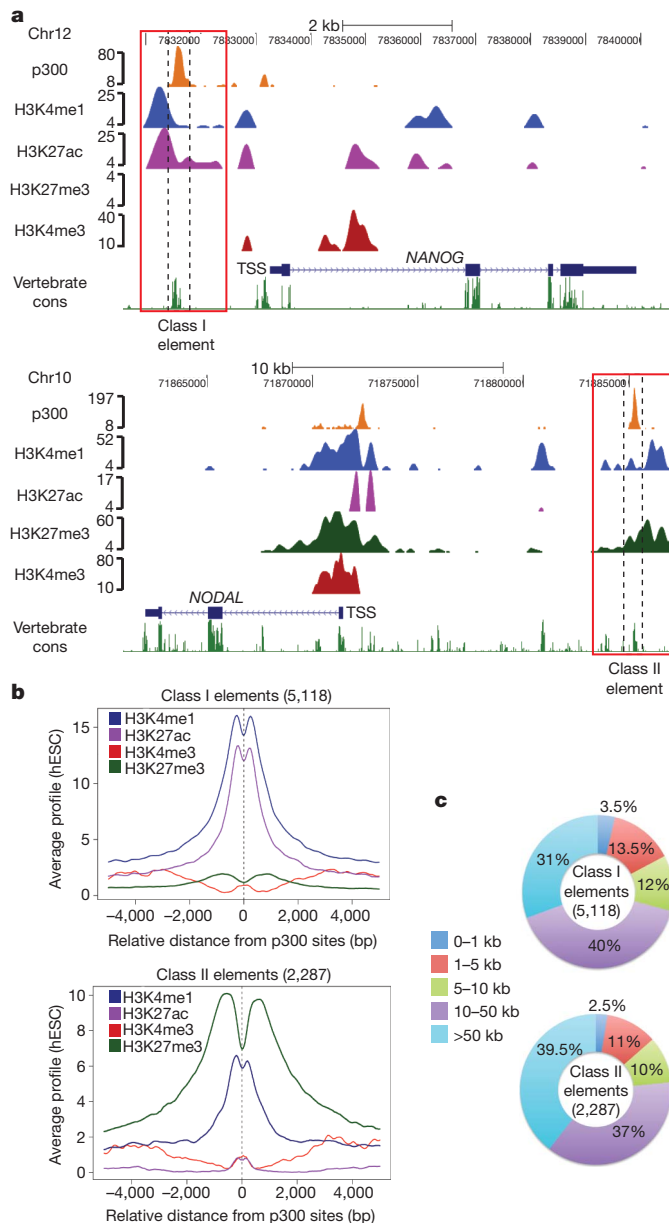


Figure 1 | Unique chromatin signatures distinguish two classes of regulatory elements in hESCs. **a**, Genome browser representations of p300, H3K4me1, H3K27ac, H3K27me3 and H3K4me3 enrichment profiles in hESCs are shown for a representative class I (for example, *NANOG*, top) and class II (for example, *NODAL*, bottom) element and its flanking regions. The peak height corresponds to normalized fold enrichments as calculated by QuEST. **b**, Average hESC ChIP-seq signal profiles were generated for the indicated histone modifications around the central position of p300-bound regions, over class I (top) and class II (bottom) elements, respectively. **c**, Class I and II elements were mapped to their closest Ensembl gene TSS and the distribution of distances between elements and TSS is shown.

To investigate whether the two classes are linked to genes of distinct functional annotations, we performed ontology analysis with the Genomic Regions Enrichment of Annotations Tool (GREAT)¹⁸ (Fig. 2d, e and Supplementary Data 2 and 3). Class I elements showed association with genes expressed in the epiblast, whose mouse homologues exhibit knockout phenotypes with defects in pre- and peri-implantation development (Fig. 2d). In contrast, class II elements are linked to genes expressed at, and essential for, gastrulation, germ-layer formation, neurulation and early somitogenesis (including *NODAL*, *EOMES*, *LEFTY2*, *EN1*, as well as *FOX*, *SOX* and *WNT* family members) (Fig. 2e). Notably, we did not observe enrichment of adult-tissue

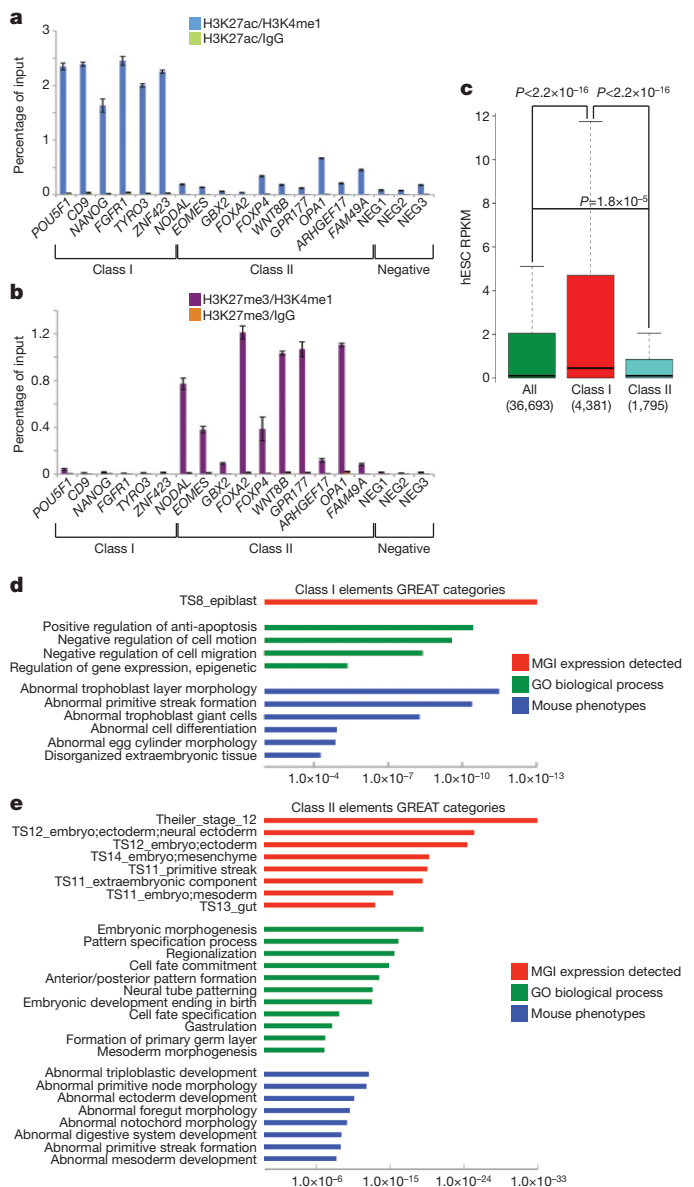


Figure 2 | Functional and molecular characterization of class I and II elements. **a**, **b**, Sequential ChIP experiments were performed from hESCs with the indicated pairs of histone modification antibodies. ChIP material was analysed by qPCR for select class I and class II elements, as well as negative control regions (NEG1–3). The y axis shows per cent input recovery; error bars represent standard deviation (s.d.) from three technical replicates. **c**, RNA-seq data set was obtained from hESC poly(A)-RNA and reads per kilobase per million mapped reads (RPKM) were calculated for all human Ensembl genes. RPKMs for all annotated genes (green) or for those closest to class I (red) or class II (blue) elements are represented as box plots. *P*-values were calculated using non-paired Wilcoxon tests. In the box plots, bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively. **d**, **e**, Functional annotation of class I (**d**) and class II (**e**) elements was performed using GREAT. The top over-represented categories belonging to three different ontologies are shown: Mouse Genome Informatics (MGI) expression detected (red) contains information on tissue- and developmental-stage-specific expression in mouse; Gene Ontology (GO) biological process (green) describes the biological processes associated with gene function; mouse phenotypes (blue) ontology contains data about mouse genotype–phenotype associations. The x axes values (in logarithmic scale) correspond to the binomial raw (uncorrected) *P*-values.

categories among class-II-linked genes, indicating no association with late enhancers.

Taken together, our results suggest that class II elements represent poised enhancers, which reveal their cell-type-dependent activity during development. One prediction from this hypothesis is that upon differentiation to a specific fate, a subset of poised enhancers linked to genes induced in this fate should acquire an active, class I signature. To test this prediction, we differentiated hESCs into neuroectodermal spheres (hNECs)¹⁹, generated p300, H3K4me1, H3K27ac and H3K27me3 profiles by ChIP-seq, and identified genomic elements that were marked by class II signature in hESCs, but acquired a strong enrichment of H3K27ac in hNECs (195 unique regions, Supplementary Data 1). Histone modification profiling over these regions showed concomitant decrease in H3K27me3 (Fig. 3a, b and Supplementary Fig. 9a) and we

refer to them hereafter as class II→I elements. Of note, a large number of the remaining class II regions (that is, those that did not acquire H3K27ac) retained H3K4me1 and H3K27me3 signature in hNECs, but showed diminished p300 occupancy (Supplementary Fig. 9b–d).

The aforementioned observations were validated by ChIP-qPCR for a representative subset of enhancers (Fig. 3c–e). We further showed that class II→I elements acquired RNA POL2 enrichment in hNECs, whereas hESC-specific active enhancers showed diminished RNA POL2 binding (Supplementary Fig. 10a). In agreement with a report documenting short bidirectional transcripts originating from enhancers²⁰, we detected an increased level of bidirectional transcription from class II→I elements upon differentiation to hNECs, whereas transcripts originating from *NANOG* and *OCT4* enhancers were downregulated (Supplementary Fig. 10b, c).

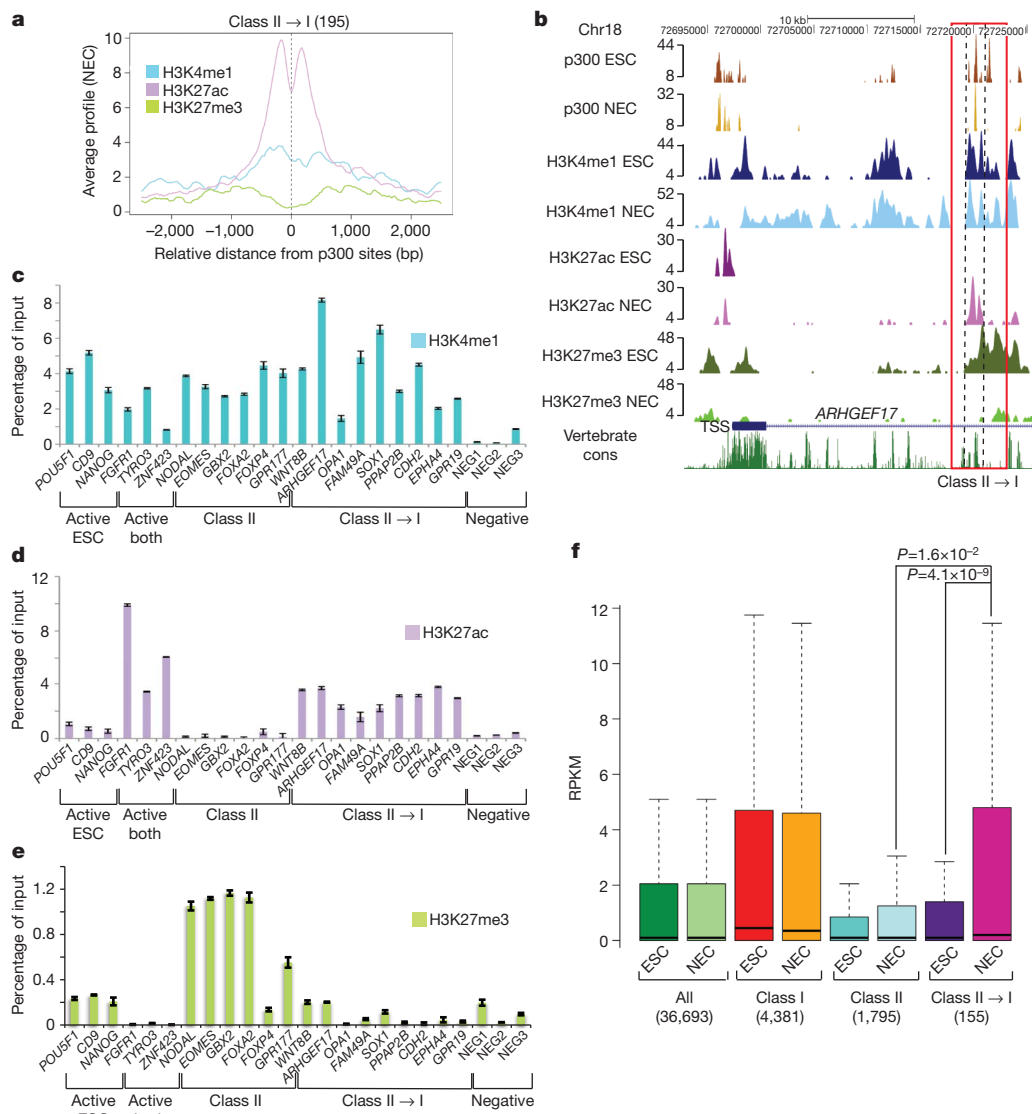


Figure 3 | A subset of class II elements acquires active enhancer chromatin signature upon neuroectodermal differentiation. **a**, Average hNEC ChIP-seq signal profiles were generated for the indicated histone modifications around the central position of those p300-bound regions (as determined in hESC) that acquired H3K27ac enrichment in hNECs (that is, class II→I elements). **b**, Genome browser representation of p300, H3K4me1, H3K27ac and H3K27me3 (in hESCs and hNECs) binding profiles at a representative class II→I element. The peak height corresponds to normalized fold enrichments as calculated by QuEST. **c–e**, ChIP-qPCR analyses from hNECs with indicated histone modification antibodies at select elements including: class I elements that were only active in hESCs (active ESC), or in both hESCs and hNECs

(active both), or class II elements that did not acquire H3K27ac in hNEC (class II), or class II→I elements. The y axis shows per cent input recovery; error bars represent s.d. from three technical replicates. ChIPs used in these qPCRs represent biological replicates of those samples used in ChIP-seq. **f**, RNA-seq data sets from hESC and hNEC poly(A)-RNA were used to calculate the RPKM for all human Ensembl genes. RPKMs in both cell types are represented as box plots for all genes (All), genes linked to class I elements, genes linked to class II elements, and genes linked to class II→I elements. *P*-values were calculated using paired (NEC class II→I versus ESC class II→I) or non-paired (NEC class II→I versus NEC class II) Wilcoxon tests.

GREAT annotation of class II→I elements showed association with genes expressed in neuroectoderm and related to abnormalities in nervous system development (Supplementary Fig. 11 and Supplementary Data 4). In agreement, hNEC RNA-seq transcriptome analysis revealed significant upregulation of class-II→I-associated genes upon differentiation, whereas expression of the remaining class-II-associated genes was persistently low (Fig. 3f). Moreover, H3K27me3 levels at class-II→I-associated TSS were diminished and H3K4me3 levels induced in hNECs as compared to hESCs, whereas modification profiles over TSS associated with the remaining class II elements were relatively unchanged (Supplementary Fig. 8b, d).

To examine if upon differentiation class II→I elements acquire the ability to drive gene expression, we infected hESCs with lentiviruses encoding a green fluorescent protein (GFP) reporter under the control of select class II→I (for example, *SOX2*, *HES1*), class I (for example, *CD9*, *JARID2*) and class II elements (for example, *EOMES*, *MYF5*) and monitored GFP fluorescence at day 1, 5 and 7 of differentiation to hNECs (Supplementary Table 1 and Supplementary Fig. 12). Class II→I reporters showed low, if any, fluorescence levels in hESCs, but were induced at day 5 of differentiation, whereas class I reporters

displayed a reverse pattern. Our results are consistent with class II elements representing poised developmental enhancers, which upon differentiation acquire, in a cell-type-dependent manner, the properties of active enhancers.

To test whether class II elements indeed function as developmental enhancers, we examined their activity during embryogenesis. Sequence conservation analysis revealed that class II elements are evolutionarily constrained and display a higher degree of conservation than class I elements (Supplementary Fig. 13a). VISTA enhancer browser search²¹ identified fourteen class II elements for which enhancer activity was previously assayed at embryonic day 11.5 of mouse development. In nine cases, highly specific expression patterns were noted (Supplementary Table 2). Interestingly, two enhancers (*WNT8B*, *CDH2*) belong to the class II→I and, in agreement, drive gene expression specifically in neuroectoderm-derived structures in the mouse (Supplementary Table 2).

Next we screened enhancer activity of a select set of class II elements using zebrafish embryo transgenic reporter assay^{22,23}. Selected elements correspond to previously uncharacterized human genomic sequences (except for *WNT8B*) that are located in proximity to genes whose zebrafish homologues have known expression patterns, although the elements themselves are generally not well conserved in the zebrafish genome (Supplementary Figs 13 and 14). GFP reporters were injected into one-cell-stage embryos and fluorescence was monitored throughout fish embryogenesis (Supplementary Fig. 15). For eight out of nine assayed class II reporters, specific and reproducible GFP patterns were observed at distinct developmental stages and anatomical locations (Fig. 4a–f, Supplementary Fig. 14 and Supplementary Table 3).

A first subgroup of assayed elements (for example, *NODAL*, *EOMES*, *LEFTY2*) drove gastrulation-specific expression at the shield, the fish equivalent of mouse primitive groove (Fig. 4a and Supplementary Fig. 14). Although none of the three tested sequences is well conserved in fish, proximal genes *NODAL*, *EOMES* and *LEFTY2* are conserved across vertebrates, with shield-specific expression pattern of zebrafish *NODAL* and *LEFTY2* homologues²⁴ (Supplementary Figs 14 and 16a). From mice to frogs, *EOMES* expression is initially restricted to the primitive groove and blastopore lip, respectively^{25,26}, but the zebrafish *EOMES* homologue is only expressed at later stages (ZFIN database, identifier ZDB-PUB-051025-1). Remarkably, the element representing a putative *EOMES* enhancer drives shield-specific expression,

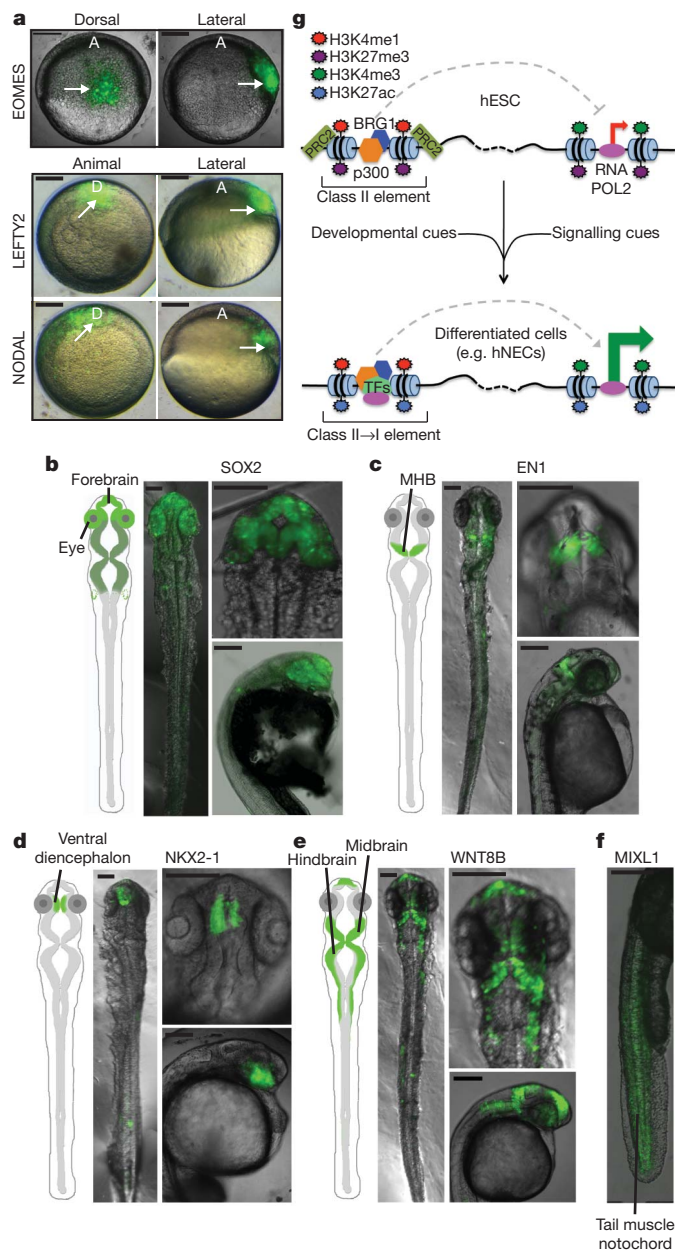


Figure 4 | Class II elements have developmental enhancer activity *in vivo*. **a**, Merged bright-field and GFP images are shown for representative shield stage zebrafish embryos injected with class II elements proximal to human *EOMES*, *LEFTY2* and *NODAL*. For the *EOMES* enhancer, dorsal (anterior to top) and lateral (shield to right) views are presented in the left and right panels, respectively. For *LEFTY2* and *NODAL*, animal pole (shield to top) and lateral (shield to right) views are presented in the left and right panels, respectively. White arrows indicate the location of the shield in each image. A, anterior; D, dorsal. Scale bar, 150 μ m. **b–f**, Merged bright-field and GFP images are shown for representative 24–28 h.p.f. zebrafish embryos injected with class II elements proximal to *SOX2* (**b**), *EN1* (**c**), *NKX2-1* (**d**), *WNT8B* (**e**) and *MIXL1* (**f**) genes. In **b–e**, schematics highlighting the relevant anatomical structures where GFP expression was reproducibly observed are shown on the left, and three images correspond, from left to right and top to bottom, to whole-embryo flattened dorsal views, dorsal anterior views and lateral anterior views, respectively. In **f**, a lateral posterior view is shown. In **b–f**, scale bar = 150 μ m. MHB, midbrain–hindbrain boundary. **g**, Proposed model for enhancer bookmarking during early embryonic development. Poised developmental enhancers (class II) are marked by a unique chromatin signature, involving occupancy of chromatin modifiers p300, BRG1 and PRC2 and nucleosomal regions marked by H3K4me1 and H3K27me3. During differentiation, appropriate developmental and signalling cues are able to rapidly transition these poised, pre-marked enhancers into an active state represented by the acquisition of H3K27ac, RNA POL2 binding, recruitment of tissue-specific transcription factors (TFs) and loss of H3K27me3, leading to the establishment of tissue-specific gene expression patterns.

indicating responsiveness of this human sequence to zebrafish gastrulation circuitry.

A second subgroup of class II reporters (for example, *SOX2*, *NKX2-1*, *EN1*, *WNT8B*, *MIXL1*) drove GFP expression at later developmental stages (24–28 h post fertilization (h.p.f.)) (Fig. 4b–f); this expression was restricted to specific anatomical structures such as the midbrain–hindbrain boundary (*EN1*)²⁷ or the ventral diencephalon/hypothalamus (*NKX2-1*)²⁸. Again, despite the low degree of sequence conservation in fish (Supplementary Fig. 13), observed GFP patterns were generally consistent with the reported expression of the putative target gene homologues^{24,29} (Supplementary Fig. 16b–d).

Importantly, specificity of our results was validated with an extensive set of control regions, including: (1) five class I elements; (2) four non-conserved genomic regions flanking select analysed class II elements; (3) four human adult tissue-specific enhancers; (4) three randomly selected intergenic non-conserved regions; (5) empty vector (Supplementary Table 4). All control regions showed only weak, diffused and nonspecific GFP patterns from 6 h.p.f. to 5 d.p.f. (Supplementary Figs 17–21). It is worth mentioning that based on our limited analysis, class I elements active in hESCs do not appear to drive pre-specification expression in zebrafish. Finally, to address whether expression patterns driven by class II elements are dynamic, we monitored several reporters (*LEFTY2*, *SOX2*, *EN1*, *NKX2-1*) throughout embryogenesis for up to 5 d.p.f. In all cases, GFP patterns were transient in nature, with fluorescence signals barely detectable after 3 d.p.f. (Supplementary Figs 17–21), further underscoring that class II regions represent dynamically regulated developmental enhancers.

We uncovered a unique chromatin signature that bookmarks early developmental enhancers in pluripotent cells, likely to prime them for a response to signalling and developmental cues (Fig. 4g). In addition to novel insights into gene regulation, our study identified a set of over 2,000 putative regulatory sequences, thereby creating an invaluable resource for lineage tracking and isolation of transient cell populations representing early steps of human development.

METHODS SUMMARY

ChIP-seq. Approximately 10^7 hESCs or hNECs were used for each ChIP experiment. Cells were crosslinked with 1% formaldehyde for 10 min at 25 °C, chromatin was sonicated and immunoprecipitated with 3–5 µg of antibody. Sequencing libraries were prepared according to Illumina protocol from: hESC and hNEC p300 ChIP, hESC BRG1 ChIP, hESC FAIRE, hESC and hNEC H3K4me3 ChIP, hESC and hNEC H3K4me1 ChIPs, hESC and hNEC H3K27me3 ChIPs, hESC and hNEC H3K27ac ChIPs, hESC and hNEC input DNAs. Libraries were sequenced using Illumina Genome Analyser and resulting sequence reads mapped by ELAND (Illumina Inc.) and analysed by QuEST 2.4 (ref. 30).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 August; accepted 25 November 2010.

Published online 15 December 2010.

1. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339**, 250–257 (2010).
2. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
3. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
4. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
5. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
6. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
7. Chan, K. K. *et al.* KLF4 and PBX1 directly regulate NANOG expression in human embryonic stem cells. *Stem Cells* **27**, 2114–2125 (2009).
8. Yeom, Y. I. *et al.* Germ-line regulatory element of Oct-4 specific for the totipotent cycle of embryonic cells. *Development* **122**, 881–894 (1996).

9. Kerppola, T. K. Polycomb group complexes—many combinations, many functions. *Trends Cell Biol.* **19**, 692–704 (2009).
10. Cockerill, P. N. *et al.* Human granulocyte-macrophage colony-stimulating factor enhancer function is associated with cooperative interactions between AP-1 and NFATp/c. *Mol. Cell. Biol.* **15**, 2071–2079 (1995).
11. Nakabayashi, H. *et al.* Functional mapping of tissue-specific elements of the human α -fetoprotein gene enhancer. *Biochem. Biophys. Res. Commun.* **318**, 773–785 (2004).
12. Itani, H. A., Liu, X., Pratt, J. H. & Sigmund, C. D. Functional characterization of polymorphisms in the kidney enhancer of the human renin gene. *Endocrinology* **148**, 1424–1430 (2007).
13. Segawa, K. *et al.* Identification of a novel distal enhancer in human adiponectin gene. *J. Endocrinol.* **200**, 107–116 (2009).
14. Mito, Y., Henikoff, J. G. & Henikoff, S. Histone replacement marks the boundaries of cis-regulatory domains. *Science* **315**, 1408–1411 (2007).
15. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genet.* **42**, 343–347 (2010).
16. Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**, 233–239 (2009).
17. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnol.* **28**, 1097–1105 (2010).
18. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28**, 495–501 (2010).
19. Bajpai, R. *et al.* Molecular stages of rapid and uniform neuralization of human embryonic stem cells. *Cell Death Differ.* **16**, 807–825 (2009).
20. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
21. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
22. Fisher, S. *et al.* Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nature Protocols* **1**, 1297–1305 (2006).
23. Navratilova, P. *et al.* Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**, 526–540 (2009).
24. Sprague, J. *et al.* The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.* **34**, D581–D585 (2006).
25. Hancock, S. N., Agulnik, S. I., Silver, L. M. & Papaioannou, V. E. Mapping and expression analysis of the mouse ortholog of *Xenopus Eomesodermin*. *Mech. Dev.* **81**, 205–208 (1999).
26. Ryan, K., Garrett, N., Mitchell, A. & Gurdon, J. B. Eomesodermin, a key early gene in *Xenopus* mesoderm differentiation. *Cell* **87**, 989–1000 (1996).
27. Danielian, P. S. & McMahon, A. P. Engrailed-1 as a target of the Wnt-1 signalling pathway in vertebrate midbrain development. *Nature* **383**, 332–334 (1996).
28. Marin, O., Baker, J., Puelles, L. & Rubenstein, J. L. Patterning of the basal telencephalon and hypothalamus is essential for guidance of cortical projections. *Development* **129**, 761–773 (2002).
29. Robb, L. *et al.* Cloning, expression analysis, and chromosomal localization of murine and human homologues of a *Xenopus* mix gene. *Dev. Dyn.* **219**, 497–504 (2000).
30. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5**, 829–834 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Wysocka laboratory members for ideas and manuscript comments; I. A. Shestopalov and J. K. Chen for sharing zebrafish resources, equipment and knowledge; T. Howes and D. M. Kingsley for the pT2HE vector; Z. Weng and A. Sidow for Illumina sequencing; and A. Valouev for discussion on ChIP-seq data analysis. This work was supported by WM Keck Foundation Distinguished Young Scholar in Biomedical Research Award and CIRRM RN1 00579-1 grant to J.W.A.R.-I. was supported by an EMBO long-term fellowship.

Author Contributions A.R.-I. conceived the project, performed and interpreted most experiments, including all genomic data analyses. R.B. established hESC culture and differentiation and performed most zebrafish imaging. T.S. generated enhancer reporter constructs, and together with S.A.B. and A.R.-I. participated in the *in vivo* enhancer screening. R.A.F. performed the RT-qPCR analysis of enhancer RNAs. J.W. contributed ideas and interpreted results. A.R.-I. and J.W. wrote the manuscript with input from all authors.

Author Information All sequencing data have been deposited in Gene Expression Omnibus (GEO) data repository under accession number GSE24447. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.W. (wysocka@stanford.edu).

METHODS

hESC culture. hESCs (H9 line, Wi-Cell) were expanded in feeder-free, serum-free medium, mTESR-1 from StemCell technologies. Cells were passaged 1:7 every 5–6 days by incubation with accutase (Invitrogen) and resultant small cell clusters (50–200 cells) were subsequently re-plated on tissue culture dishes coated overnight with growth-factor-reduced matrigel (BD Biosciences). hESC quality was regularly tested by evaluating the expression of a panel of hESC markers (for example, alkaline phosphatase, OCT4) and the capacity to differentiate into cell types derived from the three germ layers.

Neuroectoderm cell (NEC) differentiation. hESCs were differentiated into hNECs using a previously described differentiation protocol²¹. Briefly, hESCs were incubated with 2 mg ml⁻¹ collagenase. Once detached, cells were plated in NEC differentiation media: 1:1 neurobasal medium/DMEM F-12 medium (Invitrogen), 0.5× B-27 supplement minus vitamin A (50× stock, Invitrogen), 0.5× N-2 supplement (100× stock, Invitrogen), 20 ng ml⁻¹ bFGF (Peprotech), 20 ng ml⁻¹ EGF (Sigma-Aldrich), 5 µg ml⁻¹ bovine insulin (Sigma-Aldrich), 0.1 µg ml⁻¹ recombinant human NOGGIN (Peprotech), 1× Glutamax-I supplement (100× stock, Invitrogen). Cells were differentiated for 7 days, changing media every other day.

Chromatin immunoprecipitation (ChIP), sequential ChIP, FAIRE and antibodies. ChIP assays were performed from approximately 10⁷ hESCs or hNECs per experiment, according to previously described protocol with slight modifications³¹. Briefly, cells were crosslinked with 1% formaldehyde for 10 min at room temperature and formaldehyde was quenched by addition of glycine to a final concentration of 0.125 M. Chromatin was sonicated to an average size of 0.5–2 kb, using Bioruptor (Diagenode). A total of 3–5 µg of antibody was added to the sonicated chromatin and incubated overnight at 4 °C. 10% of chromatin used for each ChIP reaction was kept as input DNA. Subsequently, 75 µl of protein A or protein G Dynal magnetic beads (depending of antibody species and Ig isotype) were added to the ChIP reactions and incubated for four additional hours at 4 °C. Magnetic beads were washed and chromatin eluted, followed by reversal of the crosslinkings and DNA purification. Resultant ChIP DNA was dissolved in water.

Sequential ChIPs were performed as previously described with slight modifications³². Chromatin was prepared as described above for ChIP and after addition of the first antibody (3–5 µg) and corresponding washes, magnetic beads were resuspended in 75 µl TE/10 mM DTT. Samples were diluted 20 times with dilution buffer (1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 150 mM NaCl) and second antibody (3–5 µg) was added to each reaction. Beads were then washed, crosslinking reversed and DNA purified and dissolved in water.

For FAIRE, sonicated chromatin was prepared as for ChIP and DNA was extracted as previously described¹⁶.

All antibodies used in this study have been previously reported as suitable for ChIP: p300 (sc-585, Santa Cruz Biotechnology)⁵, BRG1 (clone JA1, a gift from G. Crabtree)³³, H3K4me1 (ab8895, Abcam)⁵, H3K27ac (ab4729, Abcam)⁵, H3K4me3 (39159, Active Motif)³⁴, H3K27me3 (39536, Active Motif)³⁵, RNA POL2 unphosphorylated (8WG16 clone, MMS-126R, Covance)³⁶, RNA POL2 ser5P (ab5131, Abcam)³⁷, RNA POL2 ser2P (ab5095, Abcam)³⁸, normal rabbit IgG (12-370, Millipore).

ChIP-qPCR. All primers used in qPCR analysis are shown in Supplementary Data 5. Primers are named after proximal putative target genes of the investigated enhancers. For each tested genomic element, two sets of primers were used, one set overlapping the peak of maximal p300 enrichment (central primers) and another set overlapping flanking regions with histone modification enrichments (flanking primers). This strategy was used because p300 peaks typically occurred within nucleosome-poor regions. qPCR analysis was performed in a Light Cycler 480II machine (Roche), using technical triplicates and ChIP-qPCR signals were calculated as percentage of input. Standard deviations were measured from the technical triplicate reactions and represented as error bars.

RT-qPCR of enhancer RNAs. To assess levels of enhancer-associated transcription, total RNA from hESCs and hNECs differentiated for 7 days was isolated using Trizol reagent followed by ethanol precipitation according to the manufacturer's protocol (Invitrogen). To remove genomic DNA contaminants, the Turbo DNA-Free kit was used following rigorous DNase treatment (two times, 30 min incubations at 37 °C). cDNA was generated from 100 ng of DNA-free RNA using the QuantiTect Reverse Transcription Kit (Qiagen) with two modifications: (1) The gDNA elimination reaction was extended for 5 min and (2) the reverse transcription elongation time was 30 min. Quantitative PCR (qPCR) primers were designed (Supplementary Data 5) to target regions surrounding the p300 peaks that defined each tested enhancer. qPCR runs and analysis were performed on the Light Cycler 480II machine (Roche). To calculate fold change between the hESCs and hNECs, the $\Delta\Delta C_t$ method was used and the 18S rRNA transcripts were used as a loading control. Standard deviations were measured from technical triplicate reactions and were represented as error bars. Biological replicate experiments for hNECs were performed and very similar results were obtained (data not shown).

ChIP-seq. Libraries were prepared from: hESC and hNEC p300 ChIP, hESC BRG1 ChIP, hESC FAIRE, hESC and hNEC H3K4me3 ChIP, hESC and hNEC H3K4me1 ChIPs, hESC and hNEC H3K27me3 ChIPs, hESC and hNEC H3K27ac ChIPs, hESC and hNEC input DNAs. ChIP-seq, FAIRE-seq and input libraries were prepared according to Illumina protocol and sequenced using Illumina Genome Analyser. All sequences were mapped by ELAND software (Illumina Inc.) and analysed by QuEST 2.4 software^{30,35}. ChIP-seq enrichment regions for the following profiled proteins were determined using the indicated settings, according to QuEST recommendations: hESC p300: KDE (kernel density estimation) bandwidth = 30, ChIP seeding fold enrichment = 30, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 3; hESC H3K4me3: KDE bandwidth = 60, ChIP seeding fold enrichment = 30, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 3; hESC H3K4me1: KDE bandwidth = 100, ChIP seeding fold enrichment = 10, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 2.5; hESC H3K27me3: KDE bandwidth = 100, ChIP seeding fold enrichment = 10, ChIP extension fold enrichment = 8, ChIP-to-background fold enrichment = 2.5; hESC and hNEC H3K27ac: KDE bandwidth = 100, ChIP seeding fold enrichment = 10, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 2.5.

For all ChIP-seq data sets, WIG files were generated with QuEST, which were subsequently used for visualization purposes and for obtaining average signal profiles.

RNA-seq. RNAs from hESCs and NECs were extracted with Trizol (Invitrogen), following the manufacturer's recommendations. 10 µg of total RNA were subjected to two rounds of oligo-dT purification using Dynal oligo-dT beads (Invitrogen). 100 ng of the purified RNA were fragmented with 10× fragmentation buffer (Ambion). Fragmented RNA was used for first-strand cDNA synthesis, using random hexamer primers (Invitrogen) and SuperScript II enzyme (Invitrogen). Second strand cDNA was obtained by adding RNaseH (Invitrogen) and DNA Pol I (New England Biolabs) to the first strand cDNA mix. The resulting double-stranded cDNA was used for Illumina library preparation as described for ChIP-seq experiments.

RNA-seq libraries were sequenced with Illumina Genome Analyser and both mapping and analysis of resulting reads were performed with DNAnexus software tools (<https://dnanexus.com>). Reads per kilobase per million mapped reads (RPKM) were calculated for all human Ensembl genes. The specificity and quality of our RNA-seq data can be visualized at several hESC- or hNEC-specific genes (Supplementary Fig. 22).

Class I and class II element selection criteria. ChIP-seq enrichment regions as determined by QuEST were used to define class I and class II elements (Supplementary Data 1). To this end, operations (intersection, subtraction, and so on) between genomic data sets were performed with GALAXY (<http://main.g2.bx.psu.edu/>) and the following selection criteria were used: class I elements (5,518 regions): genomic regions with hESC p300 enrichment (ChIP seeding fold enrichment >30), located within 2 kb of regions enriched in hESC H3K4me1 and H3K27ac (ChIP seeding fold enrichment >10 for both modifications), and, to distinguish these elements from proximal promoters, we demanded that these regions do not overlap with hESC H3K4me3 (ChIP seeding fold enrichment >30); class II elements (2,287 regions): genomic regions with hESC p300 enrichment (ChIP seeding fold enrichment >30), located within 2 kb of regions enriched in hESC H3K27me3 (ChIP seeding fold enrichment >8). These regions were further required not to overlap with hESC H3K4me3 (ChIP seeding fold enrichment >30) or hESC H3K27ac (ChIP seeding fold enrichment >10). Class II→I elements (195 regions): class II elements (as determined in hESCs) which in hNECs acquired enrichment in H3K27ac (H3K27ac ChIP seeding fold enrichment >10, within 2 kb of p300 peaks defining class II elements).

In total, we identified 11,543 regions marked by p300 and H3K4me1 in hESCs, of which 1,639 did not contain H3K27ac, H3K27me3 or H3K4me3 enrichment. A total of 3,531 regions were enriched for p300, H3K4me1 and H3K4me3 (those generally corresponded to proximal promoters).

Please note that although our definition of class II elements does not use an H3K4me1 enrichment filter, about 55% of class II regions are enriched for H3K4me1 at ChIP seeding fold enrichment >10; when lower cutoff is allowed, the overlap is significantly more substantial. Thus, the vast majority, if not all, class II elements probably contains above-background levels of H3K4me1, as exemplified by the observation that class II elements with ChIP-seq H3K4me1 levels below the seeding fold enrichment >10 cutoff are still substantially enriched for H3K4me1 when assayed by ChIP-qPCR (see Supplementary Fig. 5, for example, *CHD2*, *EPHA4*, *GPR19*, *ADRA2A*, *KLF5*, *EML1* regions).

Other sequencing data analyses. Average ChIP-seq signal profiles around the centre of p300-enriched regions were generated with the Sitepro tool, part of the Cistrome Analysis pipeline (<http://cistrome.dfci.harvard.edu/ap/>), using the corresponding WIG files generated with QuEST. Similarly, ChIP-seq signal profiles

were generated around gene TSS. For genes associated with the different classes of distal elements, each element was linked to its closest gene, based on the distance to TSS, and considering a maximum distance of 100 kb.

Average PhastCons scores profiles around the centre of p300-enriched regions were generated with the Conservation/Aggregate Datapoints tool, part of the Cistrome Analysis pipeline (<http://cistrome.dfci.harvard.edu/ap/>).

Distance between enhancers and their closest Ensembl gene TSS was calculated using PinkThing software (<http://pinkthing.cmbi.ru.nl/>) and Ensembl 52 assembly. With this information, it was possible to calculate the overall genomic distribution, based on distance to TSS, for the different enhancer groups and to assign enhancers to their closest genes.

Functional annotation of enhancers was obtained with GREAT (<http://great.stanford.edu/public/html/input.php>), using the Basal plus extension association rules and the whole human genome as background.

For RNA-seq data analysis, each enhancer was assigned to its closest gene based on distance to TSS considering a maximum distance of 100 kb, resulting in various gene groups each corresponding to an enhancer class (for example, class I, class II, class II→I). Statistical significance (*P*-values) of the difference in expression levels between different gene groups was calculated using two-sample one-sided Wilcoxon-test (R software, <http://www.r-project.org>). Paired or non-paired tests were performed when the same or different genes were compared, respectively. Box plots representing RPKM distribution were generated with R (<http://www.r-project.org>).

MRE-seq (methylation-sensitive restriction enzyme) data for hESCs was obtained from the GEO data set public repository under accession number GSM450236.

In vitro enhancer reporter assays in hESCs and hNECs. Representative class I, class II→I and class II elements (Supplementary Table 1) were cloned into a lentiviral vector (Sin-minTK-eGFP) in front of a minimal TK promoter driving GFP expression. hESC colonies were transduced with the appropriate lentiviruses and GFP fluorescence levels were subsequently monitored in undifferentiated hESCs, as well as in the course of hNEC differentiation (at day 1, 5 and 7 after induction of differentiation).

Zebrafish reporter assays. The biological relevance of the identified human enhancers was evaluated using Tol2 transposon-mediated transgenesis in zebrafish²⁴. Selected human enhancers were PCR amplified and cloned in the pT2HE vector (gift from D. M. Kingsley), upstream of the *hsp70* promoter and eGFP. Tol2 transposase was *in vitro* transcribed using mMessage mMachine Sp6 kit (Ambion), according to the manufacturer's instructions. It is worth mentioning that the *hsp70* promoter independently drives robust and stable expression in the lens after 28–38 h.p.f.³⁹. This lens signal is also observed when additional sequences are placed upstream of the minimal *hsp70* promoter, acting as a positive control for correct transgenesis. Vector DNA, with corresponding enhancers, and transposase RNA were mixed and injected in one-cell-stage zebrafish embryos as previously described. eGFP expression patterns were typically monitored at three different developmental times: 6–8 h.p.f., 10–14 h.p.f. and 24–28 h.p.f. According to ref. 24, using the described reporter assay method, 10–20% of the injected embryos are expected to display consistent and representative expression patterns. Because 50

embryos were typically injected, expression patterns were considered as representative for a given enhancer if displayed by at least 5–10 embryos within each batch (the remaining embryos typically showed a nonspecific or lack of fluorescence pattern). For those enhancers with identifiable and consistent expression patterns, a second set of injections (biological replicate) were performed for 50 additional embryos and in all cases similar results were obtained compared to the first injections.

Initial monitoring and embryo imaging were performed with a Leica M205 FA fluorescent stereoscope. High-resolution images presented in Fig. 4 were obtained with a Leica DM4500 B upright compound microscope.

Although live embryos were typically monitored and imaged, in order to obtain flat whole-embryo images, selected embryos were fixed and the yolk removed. Briefly, 24–28 h.p.f. embryos were dechorionated and transferred to 4% paraformaldehyde solution in PBS. After overnight rocking at 4 °C, fixed embryos were washed and stored in methanol at 20 °C until ready to use.

Specificity of our reporter assays was validated by assaying an extensive set of negative controls (Supplementary Table 4): (1) five class I elements; (2) four non-conserved genomic regions in proximity of four of the tested class II elements; (3) four human adult-tissue-specific enhancers that should not drive expression during early developmental stages; (4) three randomly selected intergenic non-conserved regions; (5) empty vector.

In addition, four selected class II elements were followed up to 5 days post-fertilization, together with their corresponding flanking non-conserved regions and additional negative controls. GFP patterns were monitored after 6 h.p.f., 24 h.p.f., 3 d.p.f. and 5 d.p.f. In these cases and for the class II elements, embryos showing specific patterns at the corresponding stage (for example, 6 h.p.f. for LEFTY2 and 24 h.p.f. for SOX2, EN1 and NKX2-1) were selected and their GFP patterns subsequently monitored. For the negative controls, once lens signal appeared (that is, transgenic embryos), such embryos were subsequently followed.

31. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
32. Furlan-Magaril, M., Rincon-Arango, H. & Recillas-Targa, F. Sequential chromatin immunoprecipitation protocol: ChIP-reChIP. *Methods Mol. Biol.* **543**, 253–266 (2009).
33. Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. *Proc. Natl Acad. Sci. USA* **106**, 5187–5191 (2009).
34. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).
35. Peng, J. C. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139**, 1290–1302 (2009).
36. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
37. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010).
38. Hargreaves, D. C., Horng, T. & Medzhitov, R. Control of inducible gene expression by signal-dependent transcriptional elongation. *Cell* **138**, 129–145 (2009).
39. Blechinger, S. R. *et al.* The heat-inducible zebrafish *hsp70* gene is expressed during normal lens development under non-stress conditions. *Mech. Dev.* **112**, 213–215 (2002).

lncRNAs transactivate STAUI-mediated mRNA decay by duplexing with 3' UTRs via Alu elements

Chenguang Gong^{1,2} & Lynne E. Maquat^{1,2}

Staufen 1 (STAUI)-mediated messenger RNA decay (SMD) involves the degradation of translationally active mRNAs whose 3'-untranslated regions (3' UTRs) bind to STAUI, a protein that binds to double-stranded RNA^{1,2}. Earlier studies defined the STAUI-binding site within ADP-ribosylation factor 1 (ARF1) mRNA as a 19-base-pair stem with a 100-nucleotide apex². However, we were unable to identify comparable structures in the 3' UTRs of other targets of SMD. Here we show that STAUI-binding sites can be formed by imperfect base-pairing between an Alu element in the 3' UTR of an SMD target and another Alu element in a cytoplasmic, polyadenylated long non-coding RNA (lncRNA). An individual lncRNA can downregulate a subset of SMD targets, and distinct lncRNAs can downregulate the same SMD target. These are previously unappreciated functions of non-coding RNAs and Alu elements³⁻⁵. Not all mRNAs that contain an Alu element in the 3' UTR are targeted for SMD even in the presence of a complementary lncRNA that targets other mRNAs for SMD. Most known *trans*-acting RNA effectors consist of fewer than 200 nucleotides, and these include small nucleolar RNAs and microRNAs. Our finding that the binding of STAUI to mRNAs can be transactivated by lncRNAs uncovers an unexpected strategy that cells use to recruit proteins to mRNAs and mediate the decay of these mRNAs. We name these lncRNAs half-STAUI-binding site RNAs (1/2-sbsRNAs).

Using the program mfold⁶, we failed to identify double-stranded RNA (dsRNA) structures similar to the STAUI-binding site (SBS) of ARF1 mRNA in the 3' UTRs of other SMD targets. This led us to notice that two well-characterized SMD targets—plasminogen activator inhibitor 1 (SERPINE1) mRNA and FLJ21870 mRNA (also known as ANKRD57 mRNA)^{1,2}—contain a single Alu element in their 3' UTRs. We also found that, in three independently performed microarray analyses, ~1.6% of protein-coding transcripts in HeLa cells (human epithelial cells) are upregulated at least 1.8-fold when STAUI is downregulated², and ~13% of these upregulated transcripts contain a single Alu element in their 3' UTR (Supplementary Table 1). By contrast, only ~4% of HeLa-cell protein-coding transcripts contain one or more Alu elements in their 3' UTR⁷, indicating that 3' UTR Alu elements are enriched in SMD targets relative to the bulk of cellular mRNAs.

Alu elements are the most prominent repeats in the human genome: they constitute more than 10% of the total DNA sequence in a cell and are present at up to 1.4 million copies per cell, and subfamilies of Alu elements share a 300-nucleotide consensus sequence of appreciable similarity⁸. So far, Alu elements have been documented to be *cis* effectors of protein-coding gene expression through their influence on transcription initiation or elongation, alternative splicing, adenosine to inosine (A-to-I) editing or translation initiation^{3,9}. Because non-coding RNAs (ncRNAs) that base-pair perfectly with mRNA can function in *trans* to generate endogenous short interfering RNAs (siRNAs)⁴, it seemed possible that imperfect base-pairing between the Alu element of a ncRNA and the Alu element of an mRNA 3' UTR could create an SBS, which would regulate mRNA decay. We focused on mRNAs that contain a single 3' UTR Alu element, to avoid the possibility of

intramolecular base-pairing between inverted Alu elements, which could result in A-to-I editing and nuclear retention¹⁰. Using the Antisense ncRNA Pipeline data set^{11,12}, we identified 378 lncRNAs that contain a single Alu element (Supplementary Table 2). Among them, the Alu element of the lncRNA with sequence accession number AF087999 (lncRNA_AF087999) has the potential to base-pair with the Alu element in the 3' UTR of SERPINE1 mRNA and FLJ21870 mRNA (Fig. 1a and Supplementary Fig. 1a) with ΔG values of $-151.7 \text{ kcal mol}^{-1}$ and $-182.1 \text{ kcal mol}^{-1}$, respectively (where $-151.7 \text{ kcal mol}^{-1}$ defined the most stable duplex predicted to form between SERPINE1 mRNA and any of the 378 lncRNAs) (Supplementary Table 2). This lncRNA, lncRNA_AF087999, which for reasons that follow is designated 1/2-sbsRNA1, is derived from chromosome 11.

Semiquantitative PCR with reverse transcription (RT-PCR) (Supplementary Fig. 2a) demonstrated that 1/2-sbsRNA1 is present in cytoplasmic HeLa-cell fractions but not nuclear ones and that it is polyadenylated (Supplementary Fig. 2b, c). Downregulating the cellular abundance of the two major isoforms of STAUI to <10% of normal (see, for example, Fig. 1b) did not affect either the cellular distribution or the abundance of 1/2-sbsRNA1 (Supplementary Fig. 2b). Every human tissue that was examined contained 1/2-sbsRNA1 (Supplementary Fig. 2d), and 1/2-sbsRNA1 is not a substrate for the enzymes dicer 1 (DICER1) or argonaute 2 (AGO2; also known as EIF2C2) (Supplementary Fig. 2e) and thus is distinct from the lncRNAs that generate endogenous siRNAs.

Two forms of 1/2-sbsRNA1 have been reported (NCBI sequence accession numbers AF087999 and AK094046). They differ at their 5' end but have a common Alu element and a common 3' end, which contains a putative polyadenylation signal (AUUAAA) situated 13 nucleotides upstream of a poly(A)⁺ tract. RNase protection assays confirmed the presence of one short (S) and one long (L) form of 1/2-sbsRNA1, with different 5' ends and a relative abundance in HeLa cells of 3/1 (Supplementary Fig. 3a). Primer extension (Supplementary Fig. 3b) and semiquantitative RT-PCR (Supplementary Fig. 3c) mapped the 5' end of 1/2-sbsRNA1(S) to a C nucleotide. Therefore, 1/2-sbsRNA1(S) consists of 688 nucleotides, excluding the poly(A)⁺ tract (Supplementary Fig. 3d). Whereas some transcripts that are annotated as ncRNAs may be translated⁴, data indicate that 1/2-sbsRNA1(S) is not translated (Supplementary Fig. 4).

Remarkably, in knockdown experiments, not only STAUI-directed siRNA but also 1/2-sbsRNA1-directed siRNA increased the levels of SERPINE1 and FLJ21870 mRNAs to 2–4.5-fold above normal (Fig. 1b, Supplementary Figs 5 and 6a, and Supplementary Table 3). Furthermore, experiments using the protein-synthesis inhibitor cycloheximide indicated that the 1/2-sbsRNA1-mediated reduction in SERPINE1 and FLJ21870 mRNA abundance depends on translation (Supplementary Fig. 6b), as does SMD¹³. The reduction in SERPINE1 and FLJ21870 mRNA abundance is attributable to their respective 3' UTR sequences because 1/2-sbsRNA1-directed siRNA also increased the levels of reporter (firefly luciferase, FLUC) mRNAs that contain the appropriate 3' UTR sequence (FLUC-SERPINE1 3' UTR and

¹Department of Biochemistry and Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642, USA. ²Center for RNA Biology, University of Rochester, Rochester, New York 14642, USA.

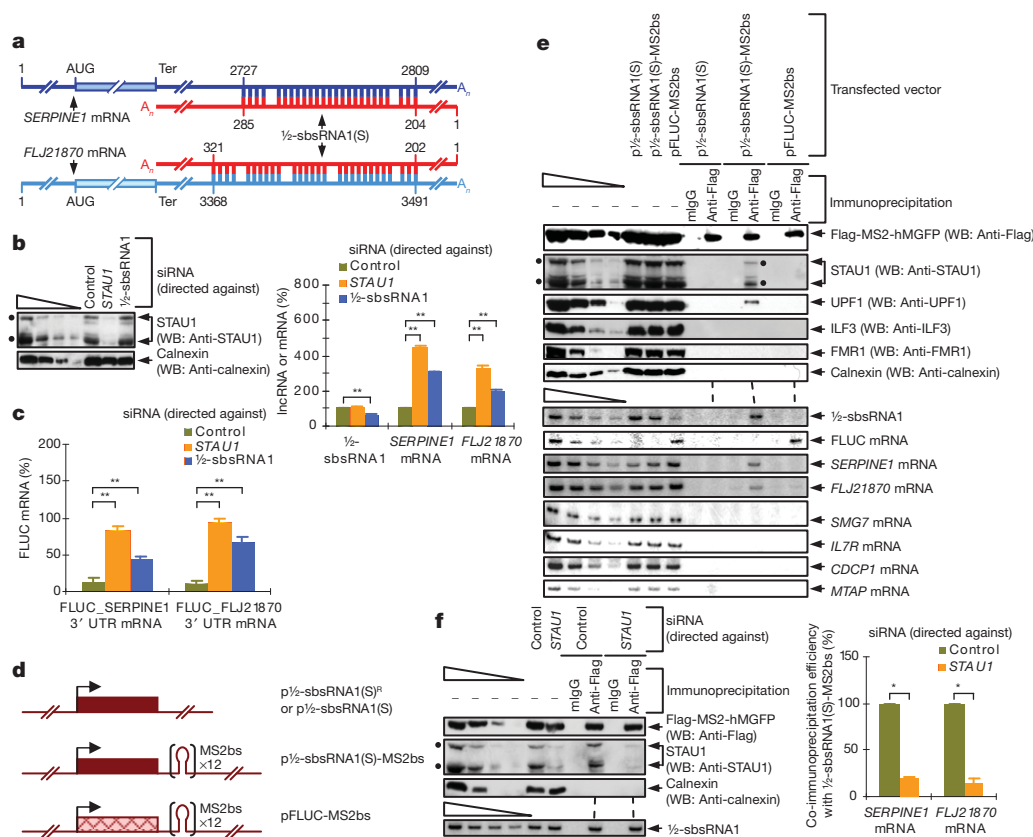


Figure 1 | 1/2-sbsRNA1 binds to, and reduces the abundance of, specific SMD targets. **a**, Predicted base-pairing between the Alu element in the 3' UTR of *SERPINE1* mRNA or *FLJ21870* mRNA and the Alu element in 1/2-sbsRNA1, where position 1 is defined as the first transcribed nucleotide of each mRNA or of 1/2-sbsRNA1(S). AUG, translation initiation codon; Ter, termination codon. **b**, Left, western blotting (WB), using the antibody designated (as anti-protein name), of lysates of HeLa cells treated with the specified siRNA. Calnexin is a loading control. Right, representation of semiquantitative RT-PCR analyses of 1/2-sbsRNA1, *SERPINE1* mRNA and *FLJ21870* mRNA from the same lysates, where the normalized level of each transcript in the presence of control siRNA is defined as 100. **c**, Representation of semiquantitative RT-PCR analyses of FLUC-*SERPINE1* 3' UTR and FLUC-*FLJ21870* 3' UTR reporter mRNAs in cells that had been transiently transfected with the specified siRNA. For each siRNA, the level of each transcript was normalized to the level of transiently expressed MUP mRNA (from the pCMV-MUP reference plasmid), where the normalized level of FLUC-No SBS mRNA was defined as 100. **d**, Diagrams of expression vectors encoding 1/2-sbsRNA1(S) or 1/2-sbsRNA1(S)^R (which differs by seven nucleotides, conferring resistance to siRNA), 1/2-sbsRNA1(S)

FLUC-*FLJ21870* 3' UTR mRNAs) relative to a reporter that does not (FLUC-No SBS mRNA) (Fig. 1c, Supplementary Fig. 5 and Supplementary Table 3). The increase in the abundance of *SERPINE1* and *FLJ21870* mRNA that is mediated by 1/2-sbsRNA1-directed siRNA was reversed by co-expression of 1/2-sbsRNA1(S)^R, which is resistant to the effects of siRNA (Supplementary Fig. 6c), arguing against siRNA-mediated off-target effects. Furthermore, 1/2-sbsRNA1-directed siRNA did not affect the expression of other FLUC reporter mRNAs that contain the 3' UTR of SMD targets that are predicted not to base-pair with 1/2-sbsRNA1 (Supplementary Fig. 7).

If 1/2-sbsRNA1 were to create an SBS by base-pairing with the 3' UTR of *SERPINE1* or *FLJ21870* mRNA, then it should be possible to co-immunoprecipitate complexes of the lncRNA and each mRNA. To test this possibility, lysates of HeLa cells that transiently expressed two plasmid DNAs were immunoprecipitated using anti-Flag antibody: the first plasmid DNA encoded 1/2-sbsRNA1(S)-MS2bs, which contains 12 copies of the MS2 coat-protein binding site (MS2bs)¹⁴

with 12 copies of MS2bs and FLUC with 12 copies of MS2bs. **e**, Western blotting (top) or semiquantitative RT-PCR (bottom) before (–) or after immunoprecipitation of lysates of formaldehyde-crosslinked HeLa cells that had been transiently transfected with pFlag-MS2-hMGFP and either the denoted 1/2-sbsRNA1(S) expression vector or pFLUC-MS2bs.

Immunoprecipitation was performed using either anti-Flag antibody or (as a control for nonspecific immunoprecipitation) mouse immunoglobulin G (mIgG). **f**, As for **e**, except for cells were treated with control or *STAU1*-directed siRNA. Western blotting (top left) and semiquantitative RT-PCR (bottom left, and right), where the co-immunoprecipitation efficiency indicates the level of each mRNA-derived product after immunoprecipitation relative to before immunoprecipitation. Each ratio in the presence of control siRNA is defined as 100%. **b**, **c**, **f**, See Supplementary Fig. 5 for phosphorimages and evaluations of semiquantitative RT-PCR data shown here as histograms. Error bars, s.e.m. *, $n = 3$, $P < 0.05$; **, $n = 6$, $P < 0.01$. **b**, **c**, **e**, Filled circles indicate the two *STAU1* isoforms. Wedges denote threefold dilutions of protein or twofold dilutions of RNA to demonstrate that analyses are semiquantitative, and dashed lines between top and bottom panels align the two panels.

upstream of the lncRNA polyadenylation signal or, as a negative control, 1/2-sbsRNA1(S) or FLUC-MS2bs mRNA (Fig. 1d); and the second plasmid DNA encoded Flag-MS2-hMGFP, which consists of the MS2 coat protein tagged with the polypeptide Flag and fused to monster green fluorescent protein (hMGFP). As expected, before immunoprecipitation, 1/2-sbsRNA1(S), as well as 1/2-sbsRNA1(S)-MS2bs, decreased the abundance of *SERPINE1* and *FLJ21870* mRNA but not other SMD targets that encode the interleukin-7 receptor (IL-7R), CUB-domain-containing protein 1 (CDCP1) or methylthioadenosine phosphorylase (MTAP) (Fig. 1e).

In support of our hypothesis that 1/2-sbsRNA1 creates an SBS with partially complementary mRNA sequences, in lysates of cells expressing 1/2-sbsRNA1(S)-MS2bs, the anti-Flag-antibody-mediated immunoprecipitation of Flag-MS2-hMGFP bound to 1/2-sbsRNA1(S)-MS2bs co-immunoprecipitated endogenous *STAU1* and *SERPINE1* and *FLJ21870* mRNA, as well as UPF1, a factor involved in SMD (Fig. 1e). By contrast, irrelevant proteins, such as calnexin, the dsRNA-binding protein ILF3

(ref. 15), the single-stranded RNA-binding protein FMR1 (ref. 16), and mRNAs that are predicted not to base-pair with 1/2-sbsRNA1, such as those encoding SMG7, IL-7R, CDCP1 and MTAP, were not co-immunoprecipitated (Fig. 1e). *STAU1*-directed siRNA reduced the co-immunoprecipitation of *SERPINE1* mRNA or *FLJ21870* mRNA with 1/2-sbsRNA1(S)-MS2bs to ~19% of normal or ~15% of normal, respectively (Fig. 1f and Supplementary Fig. 5), indicating that STAU1 stabilizes the duplex that is formed between *SERPINE1* or *FLJ21870* mRNA and 1/2-sbsRNA1.

As additional evidence that 1/2-sbsRNA1 creates an SBS by base-pairing with the 3' UTR of *SERPINE1* or *FLJ21870* mRNA, only STAU1 (tagged with the polypeptide HA₃) but not ILF3 or FMR1 co-immunoprecipitated with 1/2-sbsRNA1 (Supplementary Fig. 8).

To determine whether 1/2-sbsRNA1 is required for the co-immunoprecipitation of STAU1 with *SERPINE1* or *FLJ21870* mRNA, HeLa cells that transiently expressed STAU1-HA₃ and control siRNA or 1/2-sbsRNA1-directed siRNA in the presence or absence of 1/2-sbsRNA1(S)^R were immunoprecipitated using anti-HA antibodies. Compared with control siRNA, 1/2-sbsRNA1-directed siRNA (which reduced the level of 1/2-sbsRNA1 to ~50% of normal) reduced by about twofold the co-immunoprecipitation of STAU1-HA₃ with *SERPINE1* or *FLJ21870* mRNA (Fig. 2a and Supplementary Fig. 5). By contrast, restoring the level of 1/2-sbsRNA1 to ~100% of normal by expressing 1/2-sbsRNA1-directed siRNA together with 1/2-sbsRNA1(S)^R restored the co-immunoprecipitation of STAU1-HA₃ with *SERPINE1* or *FLJ21870* mRNA to near normal (Fig. 2a and Supplementary Fig. 5). As expected, the level of *IL7R* mRNA, which binds to STAU1 (ref. 2) but does not contain sequences complementary to 1/2-sbsRNA1, was unaffected by any of these conditions

either before or after immunoprecipitation (Fig. 2a and Supplementary Fig. 5).

We conclude that the SMD of *SERPINE1* and *FLJ21870* mRNA involves base-pairing between their 3' UTR Alu element and the Alu element in 1/2-sbsRNA1. Base-pairing creates an SBS that is stabilized by STAU1. Furthermore, the level of STAU1, and thus the efficiency of SMD, does not alter the level of 1/2-sbsRNA1. Our finding that downregulating *SERPINE1* mRNA to 50% of normal and *FLJ21870* mRNA to 25% of normal failed to detectably decrease the co-immunoprecipitation of STAU1-HA₃ with 1/2-sbsRNA1 (Supplementary Fig. 9) indicates that 1/2-sbsRNA1 may bind to more than just *SERPINE1* and *FLJ21870* mRNAs to recruit STAU1 if not trigger SMD.

The presence of UPF1 in the anti-Flag-antibody-mediated immunoprecipitation of Flag-MS2-hMGFP (Fig. 1e) is consistent with the idea that STAU1 that is bound to a 1/2-sbsRNA1-created SBS associates with UPF1, analogously to how STAU1 that is bound to the *ARF1* mRNA SBS associates with UPF1 (refs 2, 13). Furthermore, downregulating UPF1, like downregulating STAU1, increases the abundance of *SERPINE1* mRNA, *FLJ21870* mRNA and FLUC-*SERPINE1* 3' UTR mRNA by increasing mRNA half-life^{1,2}. To test UPF1 function in conjunction with 1/2-sbsRNA1, we analysed the effects of various siRNAs on the production of FLUC-*SERPINE1* 3' UTR mRNA with different versions of the 3' UTR: an intact 3' UTR; a 3' UTR that precisely lacked the region that is partially complementary to 1/2-sbsRNA1 (the binding site (BS) region); and a 3' UTR that contained only this region (Fig. 2b). Relative to control siRNA, *STAU1*-directed siRNA, *UPF1*-directed siRNA or 1/2-sbsRNA1-directed siRNA did not affect the level of FLUC-*SERPINE1* 3' UTR mRNA that lacked the 1/2-sbsRNA1-BS

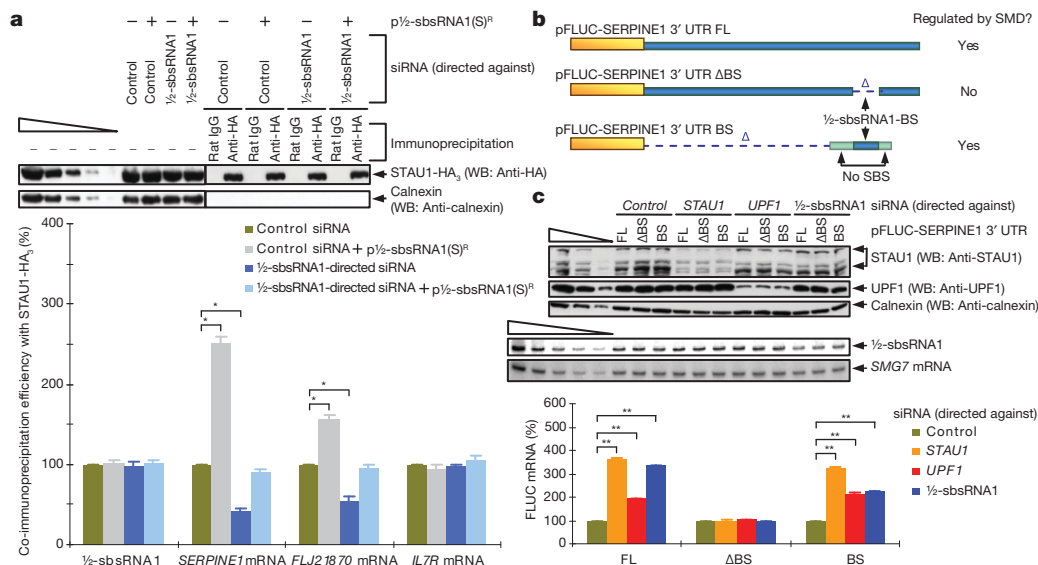


Figure 2 | 1/2-sbsRNA1 co-immunoprecipitates with STAU1 and is required for STAU1 binding to specific SMD targets. **a**, Western blotting (top) or semiquantitative RT-PCR (bottom) of lysates of formaldehyde-crosslinked HeLa cells that had been transiently transfected with the specified siRNA and either empty vector (–) or p1/2-sbsRNA1(S)^R (+) before or after immunoprecipitation with anti-HA antibody or rat IgG. After immunoprecipitation, each sample was spiked with *in vitro*-synthesized *Escherichia coli lacZ* mRNA as a loading control. The co-immunoprecipitation efficiency provides the level of each mRNA semiquantitative RT-PCR product after immunoprecipitation relative to before immunoprecipitation, where each ratio in the presence of control siRNA is defined as 100%. **b**, Diagram of pFLUC-*SERPINE1* 3' UTR FL, which contains the full-length (FL) *SERPINE1* 3' UTR, and two 3' UTR deletion variants: pFLUC-*SERPINE1* 3' UTR BS contains only the 1/2-sbsRNA1-binding site (BS), and pFLUC-*SERPINE1* 3' UTR ΔBS contains the entire 3' UTR except for the BS. FLUC sequences are shown in yellow, *SERPINE1* 3' UTR sequences in blue, and the 3' UTR of FLUC-No SBS, which does not bind to STAU1, in green; Δ indicates deleted sequence. The 5'-most green box ensures that ribosomes translating to the FLUC termination codon do not displace STAU1 that has been recruited to the 1/2-sbsRNA1-BS (which is 86 nucleotides, as shown in Supplementary Fig. 1a). **c**, Western blotting (top) and semiquantitative RT-PCR (centre and bottom) of lysates of HeLa cells that had been transiently transfected with the noted pFLUC-*SERPINE1* 3' UTR construct and the pCMV-MUP reference plasmid. Bottom, the normalized level of each FLUC mRNA in the presence of control siRNA is defined as 100%. **a**, **c**, See Supplementary Fig. 5 for phosphorimages and evaluation of semiquantitative RT-PCR data shown here as histograms. **a**–**c**, Error bars, s.e.m. *, $n = 3$, $P < 0.05$; **, $n = 6$, $P < 0.01$.

(Fig. 2c and Supplementary Fig. 5). However, each of these siRNAs increased the levels of FLUC-SERPINE1 3' UTR mRNA and FLUC mRNA that contained only the 1/2-sbsRNA1-BS (Fig. 2c and Supplementary Fig. 5). We conclude that, as indicated by its name, 1/2-sbsRNA1 base-pairs with the 3' UTR of *SERPINE1* mRNA and, by analogy, *FLJ21870* mRNA to recruit STAU1 and its binding partner UPF1 in a way that triggers a reduction in mRNA abundance. Consistent with previous studies of SMD^{2,13}, the STAU1- and 1/2-sbsRNA1-mediated reduction in mRNA abundance is due to a decrease in mRNA half-life (Supplementary Fig. 10). With regard to function, scrape-injury repair assays showed that 1/2-sbsRNA1 contributes towards reducing cell migration by targeting *SERPINE1* and RAB11-family-interacting protein 1 (*RAB11FIP1*) mRNA for SMD (Supplementary Fig. 11).

Characterizing seven other lncRNAs that contain a single Alu element and consist of <1,000 nucleotides (Supplementary Table 2) confirmed that they too are largely cytoplasmic and polyadenylated (Supplementary Fig. 2b, c and data not shown), and they have the potential to base-pair with the single Alu element in at least one mRNA 3' UTR (Fig. 3a, Supplementary Fig. 1b–d, Supplementary Table 2 and data not shown). Individually downregulating three of these lncRNAs—lncRNA_BC058830 (1/2-sbsRNA2), lncRNA_AF075069 (1/2-sbsRNA3) or lncRNA_BC009800 (1/2-sbsRNA4)—upregulated those tested mRNAs that contain a partially complementary Alu element and are upregulated on STAU1 or UPF1 downregulation; downregulation of each lncRNA failed to upregulate mRNAs that lack a partially complementary Alu element (Fig. 3b, Supplementary Fig. 5 and data not shown). Whereas 1/2-sbsRNA2 targeted the 3' UTR Alu

element of *CDCP1* mRNA ($\Delta G = -153.7 \text{ kcal mol}^{-1}$) (Fig. 3b, Supplementary Fig. 5 and Supplementary Table 2), 1/2-sbsRNA3 and 1/2-sbsRNA4 targeted the 3' UTR Alu element of *MTAP* mRNA ($\Delta G = -203.1$ and $-264.2 \text{ kcal mol}^{-1}$, respectively) (Fig. 3b, Supplementary Fig. 5 and Supplementary Table 2). Furthermore, none of the three lncRNAs downregulated *SERPINE1* mRNA ($\Delta G = 0$, -66.4 and $-108.2 \text{ kcal mol}^{-1}$ for 1/2-sbsRNA2, 1/2-sbsRNA3 and 1/2-sbsRNA4, respectively) (Fig. 3b, Supplementary Fig. 5 and Supplementary Table 2), but two of them downregulated *FLJ21870* mRNA about twofold ($\Delta G = -261.9$ and $-444.2 \text{ kcal mol}^{-1}$ for 1/2-sbsRNA3 and 1/2-sbsRNA4, respectively) (Fig. 3b, Supplementary Fig. 5 and Supplementary Table 2).

These findings illustrate the potentially complex network of regulatory events that are controlled by lncRNA–mRNA duplexes that bind to STAU1, a network that is reminiscent of the web of regulatory mechanisms that are mediated by microRNAs¹⁷. Notably, both *CDCP1* mRNA and *MTAP* mRNA were upregulated at least twofold on STAU1 downregulation in experiments reported here (Fig. 3b), and indeed *CDCP1* mRNA is among those mRNAs that were upregulated minimally (1.8 fold) on STAU1 downregulation² (Supplementary Table 1). However, because *MTAP* mRNA was upregulated only ~1.5 fold², it is not included in Supplementary Table 1. Thus, Supplementary Table 1 must be considered to provide only a partial list of mRNAs that are modulated by one or more 1/2-sbsRNAs. Conceivably, the degree of modulation could vary in different cell types (Supplementary Fig. 2d) or developmental stages depending on the abundance of the 1/2-sbsRNA(S) and on proteins that inhibit or enhance base-pairing.

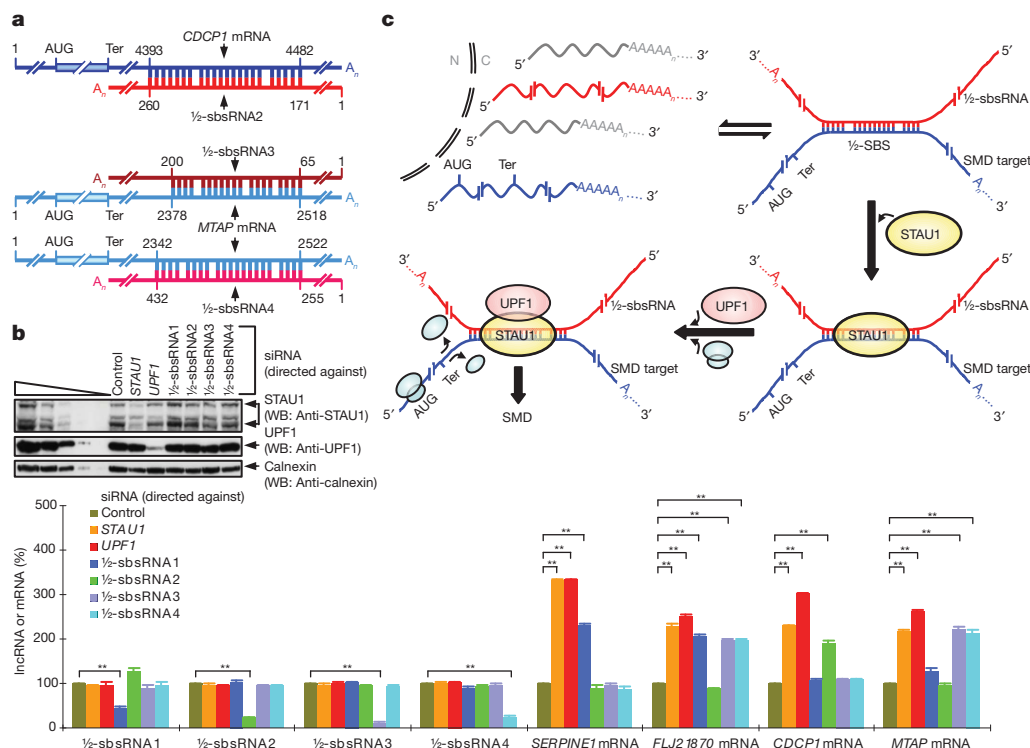


Figure 3 | Evidence that 1/2-sbsRNA2, 1/2-sbsRNA3 and 1/2-sbsRNA4 base-pair with particular mRNA 3' UTRs and decrease mRNA abundance, as do STAU1 and UPF1. **a**, Predicted base-pairing between the 3' UTR Alu element of *CDCP1* mRNA (NCBI Nucleotide accession number NM_022842) and 1/2-sbsRNA2, or *MTAP* mRNA (accession number NM_002451) and 1/2-sbsRNA3 as well as 1/2-sbsRNA4, where position 1 is defined as the first nucleotide listed in the NCBI Nucleotide database for each mRNA or lncRNA. **b**, Essentially as in Fig. 1b. See Supplementary Fig. 5 for phosphorimages and evaluation of semiquantitative RT-PCR data shown here as histograms. Error bars, s.e.m. **, $n = 6$, $P < 0.01$. **c**, Model for how an Alu-element-containing

1/2-sbsRNA that is polyadenylated and largely cytoplasmic (red) base-pairs with a partially complementary Alu element, that is, a half-STAU1 binding site (1/2-SBS), within the 3' UTR of a particular mRNA (blue) to trigger SMD. Base-pairing forms a functional SBS. The STAU1-bound SBS triggers SMD in a UPF1-dependent mechanism when translation terminates sufficiently upstream of the SBS so that translating ribosomes (blue ovals) do not remove bound STAU1. The 1/2-sbsRNA is not destroyed in the process. C, cytoplasm; N, nucleus; Ter, termination codon (which is generally, but not necessarily, a normal termination codon).

It is important to note that ΔG values are not in themselves absolute predictors of SBS function. For example, although 1/2-sbsRNA2 is predicted to base-pair with the 3' UTR Alu element of BAG5 mRNA (with $\Delta G = -416 \text{ kcal mol}^{-1}$), BAG5 mRNA is not targeted for SMD in HeLa cells (Supplementary Fig. 12). The 3' UTR Alu element of BAG5 mRNA may be physically inaccessible to base-pairing with 1/2-sbsRNA2. Nevertheless, base-pairing itself may not be sufficient for SBS function because converting the 100-nucleotide apex of the intramolecular ARF1 mRNA SBS to a 4-nucleotide loop that is predicted not to disrupt the adjacent 19-base-pair stem of the ARF1 mRNA SBS reduces STAU1 binding *in vivo* by 50% (ref. 2).

Here we report an unforeseen role for some of the lncRNAs that contain Alu elements: the creation of SBSs by intermolecular base-pairing with an Alu element in the 3' UTR of one or more mRNAs. We conclude that SBSs can form either through intramolecular base-pairing, as exemplified by the ARF1 mRNA SBS, or intermolecular base-pairing between a 1/2-SBS in an mRNA 3' UTR and a complementary 1/2-sbsRNA in the form of a largely cytoplasmic lncRNA (Fig. 3c).

There are estimated to be tens of thousands of human ncRNAs that have little or no ability to direct protein synthesis and that are distinct from ribosomal RNAs, transfer RNAs, small nuclear RNAs, small nucleolar RNAs, siRNAs and microRNAs¹⁸. Thus, the model in which partially complementary ncRNA–mRNA duplexes can form SBSs may extend to the creation of binding sites for other dsRNA-binding proteins. Because only 23% of lncRNAs were found to contain one or more Alu elements, it is also possible that lncRNA–mRNA duplexes that do not involve Alu elements could increase the number of ncRNAs that regulate gene expression by SMD or a different dsRNA-binding-protein-dependent pathway.

METHODS SUMMARY

A Perl program, Alu_Mask, was written and used together with the program RepeatMasker (<http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>) to define Alu elements. A Perl program, RNA_RNA_anneal, which uses a recursive algorithm^{19,20}, was generated to predict intermolecular duplexes between Alu elements in lncRNAs and proven or putative SMD targets. Duplexes were then validated using the program RNAstructure 4.6 (<http://rna.urmc.rochester.edu/RNAstructure.html>), which provides folding free energy changes (ΔG). Human cells (HeLa or HaCaT) were transiently transfected with the specified plasmids or siRNAs as previously described¹. For mRNA half-life measurements, Tet-Off HeLa cells (Clontech) were used. If the cells had been crosslinked with formaldehyde, cells were sonicated six times for 30 s to facilitate lysis, and crosslinks were reversed by heating at 65 °C for 45 min after immunoprecipitation. Immunoprecipitation was performed as previously described¹. Protein was purified, and western blotting was performed as previously described¹. RNA was purified from total, nuclear or cytoplasmic cell fractions or immunoprecipitated from total-cell lysates as previously reported¹. Poly(A)⁺ RNA was extracted from total-cell RNA by using the Oligotex mRNA Mini Kit (Qiagen). Semiquantitative RT–PCR and quantitative real-time RT–PCR were carried out as previously described¹, except in certain cases in which RT was primed using oligo(dT)₁₈ rather than random hexamers. For RNase protection assays, the RPA III Ribonuclease Protection Assay Kit (Ambion) was used, together with uniformly labelled RNA probes that were generated by transcribing linearized pcDNA3.1/Zeo(+)-Chr11_66193000–66191383 (which contains 1/2-sbsRNA1(S) and upstream and downstream flanking sequences) *in vitro* using [α -³²P]UTP (PerkinElmer) and the MAXIscript T7 Kit (Ambion). Cells were visualized using an Eclipse TE2000-U inverted fluorescence microscope (Nikon), and a 480-nm excitation wavelength was used for phase contrast microscopy. Images were captured using TILLvisiON software (TILL Photonics). Scrape-injury repair assays were performed essentially as previously published^{21,22}. All data were derived from at least three independently performed experiments that did not vary by more than the amount shown, and *P* values for all semiquantitative RT–PCR results were <0.05. All *P* values were determined by one-tailed *t*-tests.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 June; accepted 25 November 2010.

- Gong, C., Kim, Y. K., Woeller, C. F., Tang, Y. & Maquat, L. E. SMD and NMD are competitive pathways that contribute to myogenesis: effects on PAX3 and myogenin mRNAs. *Genes Dev.* **23**, 54–66 (2009).
- Kim, Y. K. *et al.* Stau1 regulates diverse classes of mammalian transcripts. *EMBO J.* **26**, 2670–2681 (2007).
- Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.* **10**, 691–703 (2009).
- Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* **23**, 1494–1504 (2009).
- Walters, R. D., Kugel, J. F. & Goodrich, J. A. InvAluable junk: the cellular impact and function of Alu and B2 RNAs. *IUBMB Life* **61**, 831–837 (2009).
- Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
- Yulug, I. G., Yulug, A. & Fisher, E. M. The frequency and position of Alu repeats in cDNAs, as determined by database searching. *Genomics* **27**, 544–548 (1995).
- Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379 (2002).
- Hasler, J. & Strub, K. Alu elements as regulators of gene expression. *Nucleic Acids Res.* **34**, 5491–5497 (2006).
- Chen, L. L., DeCervo, J. N. & Carmichael, G. G. Alu element-mediated gene silencing. *EMBO J.* **27**, 1694–1705 (2008).
- Pang, K. C. *et al.* RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.* **35**, D178–D182 (2007).
- Engström, P. G. *et al.* Complex loci in human and mouse genomes. *PLoS Genet.* **2**, e47 (2006).
- Kim, Y. K., Furic, L., Desgroseillers, L. & Maquat, L. E. Mammalian Stau1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell* **120**, 195–208 (2005).
- Kim, H. H. *et al.* HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev.* **23**, 1743–1748 (2009).
- Kuwano, Y. *et al.* NF90 selectively represses the translation of target mRNAs bearing an AU-rich signature motif. *Nucleic Acids Res.* **38**, 225–238 (2010).
- Ashley, C. T. Jr, Wilkinson, K. D., Reines, D. & Warren, S. T. FMR1 protein: conserved RNP family domains and selective RNA binding. *Science* **262**, 563–566 (1993).
- Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
- Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
- Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).
- Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
- Liang, C. C., Park, A. Y. & Guan, J. L. *In vitro* scratch assay: a convenient and inexpensive method for analysis of cell migration *in vitro*. *Nature Protocols* **2**, 329–333 (2007).
- Providence, K. M. *et al.* SERPINE1 (PAI-1) is deposited into keratinocyte migration 'trails' and required for optimal monolayer wound repair. *Arch. Dermatol. Res.* **300**, 303–310 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Mathews and A. Grossfield for the use of computer clusters, D. Mathews for access to the program Structure 5.0, K. Nerhrke for fluorescence microscope time, S. de Lucas and J. Ortiz for anti-STAU1 antibodies, M. Gorospe for pcDNA3-MS2bsX12, S. Higgins and P. Higgins for HaCaT cells and advice on the scrape-injury repair assay, J. Wang for the initial BAG5 mRNA assays, and O. Isken, M. Gleghorn and D. Mathews for comments on the manuscript. This work was supported by the National Institutes of Health (GM074593 to L.E.M.) and an Elon Huntington Hooker Graduate Student Fellowship (C.G.).

Author Contributions C.G. wrote the Perl programs and performed the bioinformatics analyses and wet-bench experiments. C.G. and L.E.M. analysed the computational data, designed the wet-bench experiments, analysed the resultant data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.E.M. (lynne_maquat@urmc.rochester.edu).

METHODS

Computational analyses. A Perl program, *Alu_Mask*, was designed to define Alu elements within known and putative SMD targets and ncRNAs, on the basis of results obtained using the program RepeatMasker (<http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>).

A Perl program, *RNA_RNA_anneal*, was developed to predict Alu-element base-pairing between lncRNA_AF087999 (1/2-sbsRNA1) and the *SERPINE1* or *FLJ21870* mRNA 3' UTR, between lncRNA_BC058830 (1/2-sbsRNA2) and the *CDCP1* mRNA 3' UTR, and between lncRNA_AF075069 (1/2-sbsRNA3) or lncRNA_BC009800 (1/2-sbsRNA4) and the *MTAP* or *FLJ21870* mRNA 3' UTR. Potential duplexes were fixed using the region that was predicted to be the most stably and perfectly base-paired and then expanded in both directions, allowing bulges or loops of up to 10 nucleotides, until base-pairing was no longer predicted. Briefly, *RNA_RNA_anneal* uses a recursive algorithm that predicts the most stable base pairs and their folding free-energy change (ΔG) based on thermodynamic data^{19,20} that were extracted from the program RNAstructure 4.6 (<http://rna.urmc.rochester.edu/rnastructure.html>). Duplexes between other ncRNAs and mRNA 3' UTRs were likewise predicted using this approach. All data from *RNA_RNA_anneal* were validated using the program RNAstructure 4.6, which also provides ΔG values.

Notably, to follow up our finding that ~13% of the ~1.6% of HeLa-cell protein-coding transcripts that are upregulated at least 1.8-fold on STAU1 downregulation^{1,2} contain a single Alu element, a random resampling of 1.6% of total-cell mRNAs (NCBI RefSeq) 10,000 times showed that the presence of one or more Alu elements in the 3' UTRs of potential SMD targets (Supplementary Table 1) was enriched ~3.58-fold ($P < 0.001$).

Perl program codes are available for downloading from http://dbb.urmc.rochester.edu/labs/maquat/maquat_lab.htm.

Plasmid constructions. To construct pcDNA3.1/Zeo(+)-Chr11_66193000–66191383, HeLa-cell genomic DNA was purified using DNeasy Blood & Tissue Kit (Qiagen) and amplified by PCR using the primer pair 5'-GATGCTCGAGTGGCATTGGCTTTACCACCTATG-3' (sense) and 5'-GTCAGGATCCTGCCTCAAGTCAAAGCACAACCTG-3' (antisense), where the underlined nucleotides specify a XhoI or BamHI site, respectively. The resultant PCR product was cleaved with XhoI and BamHI and inserted into XhoI- and BamHI-cleaved pcDNA3.1/Zeo(+) vector (Invitrogen).

To generate p1/2-sbsRNA1(S) or p1/2-sbsRNA1(S)-MS2bs, pcDNA3.1/Zeo(+)-Chr11_66193000–66191383 was amplified using the primer pair 5'-GAGTCAAAGCTTAAAGGAGAGACAGTCTCACTCTG-3' (sense) and 5'-GTCAGCGGCCGCCAGTTGTGAAGCATATTTGGGTTAC-3' (antisense) or 5'-GTCAGGATCCAGTTGTGAAGCATATTTGGGTTAC-3' (antisense), respectively, where underlined nucleotides denote a HindIII, NotI or BamHI site, respectively. The resultant PCR products were cleaved with HindIII and either NotI or BamHI, respectively, and inserted into HindIII- and NotI-cleaved or HindIII- and BamHI-cleaved pcDNA3-MS2bs¹⁴.

Overlap extension PCR was used to construct p1/2-sbsRNA1(S)^R. Two rounds of site-directed mutagenesis were performed using p1/2-sbsRNA1(S) and the following primer pairs: first round, 5'-GATATTCATTACTAACCCTGAACCCATACAGTTCAGCTTACCCTACAGTACTTCT-3' (sense) and 5'-AGAA GTACTGTAGTGGTAAAGCTGAACCTGTATGGGTTTCAGGGTTAGTAATGATATC-3' (antisense); and second round, 5'-CCTGAACCCATACAGTTCAGTACAGAACTACAGTACTTCTGTAGT-3' (sense) and 5'-ACTACAGAACTACTGTAGTCTGAGCTGAACCTGTATGGGTTTCAGG-3' (antisense), where mutagenic nucleotides are underlined.

To generate pFLUC-MS2bs, pFLUC¹³ was amplified by PCR using the primer pair 5'-GAGTCAAAGCTTATGGAAGACGCCAAAAACATAAAGAAAGGC-3' (sense) and 5'-GTCAGGATCCTTACAATTGGACITTCGCCCTTCTTG GC-3' (antisense), where underlined nucleotides specify a HindIII or BamHI site. The resultant PCR product was digested with HindIII and BamHI and inserted into HindIII- and BamHI-cleaved pcDNA3-MS2bs.

To construct pFlag-MS2-hMGFP, pMS2-HA¹³ was amplified using the primer pair 5'-GATGGCTAGCCGCATGGACTACAAAGACGATGACGACAAGG GATCCGCTTCTAAGTTTACTCAGTTCG-3' (sense) and 5'-GTCAGATATC GTAGATGCCGGAGTTTGTCTGCG-3' (antisense), where underlined nucleotides specify an NheI or EcoRV site. The resultant PCR product was digested using NheI and EcoRV and inserted into NheI- and EcoRV-cleaved pMGFP vector (Promega).

To create p1/2-sbsRNA1(S)-hMGFP, pMGFP was amplified using the primer pair 5'-GATGCCTAGGGGCGTGATCAAGCCCGACATG-3' (sense) and 5'-GTCACTAGGGCGGCGCTGGCGGGGTAGTCC-3' (antisense), where underlined nucleotides identify the AvrII site. The resultant PCR product was digested with AvrII and inserted into the AvrII site of p1/2-sbsRNA1(S).

To construct pFLUC-FLJ21870 3' UTR, two fragments of the *FLJ21870* mRNA 3' UTR were amplified using HeLa-cell genomic DNA and the primer pairs 5'-GATGCTAGAGTGATCAACTTCGCCAACAAACACCAG-3' (sense) and 5'-CAGAAGGCTAGCCCCAAGAGAAC-3' (antisense), and 5'-CTCTTCGG GCTAGCCTTCTGG-3' (sense) and 5'-GTCAGGGCCCCGAGACAGAGTCTC CGTTGCCC-3' (antisense), where underlined nucleotides denote an XbaI, NheI, NheI or ApaI site, respectively. The resultant PCR fragments were digested using NheI and either XbaI or ApaI, and inserted simultaneously into pFLUC-SERPINE1 3' UTR⁵ that had been digested with XbaI and ApaI.

To create pFLUC-SERPINE1 3' UTR Δ (1/2-sbsRNA1-BS), two regions of the *SERPINE1* mRNA 3' UTR were amplified using pFLUC-SERPINE1 3' UTR and the primer pairs 5'-GAGTCAAAGCTTGGCATTCCGGTACTGTTGG-3' (sense) and 5'-CATCCATCTTTGTGCCCACCC-3' (antisense), and 5'-TCTTTAAA AATATATATATTTTAAATATAC-3' (sense) and 5'-TAGAAGGCACAGTCG AGG-3' (antisense), where underlined nucleotides denote a HindIII site. The resultant PCR fragments were phosphorylated using T4 polynucleotide kinase, digested with HindIII or ApaI (which binds upstream of where the antisense primer anneals), respectively, and inserted simultaneously into pFLUC-SERPINE1 3' UTR that had been digested with HindIII and ApaI.

To generate pFLUC-SERPINE1 1/2-sbsRNA1-BS, 1/2-sbsRNA1-BS was amplified using pFLUC-SERPINE1 3' UTR and the primer pair 5'-GATGTTTA AATAATGCACCTTTGGGAGGCCAAGG-3' (sense) and 5'-GATGTTTAAAG ACGGGGCTTGGTATGTGTC-3' (antisense), where underlined nucleotides denote a DraI site. The resultant PCR product was then digested with DraI. Meanwhile, pFLUC-No SBS was digested with HindIII and ApaI, and the released FLUC-No SBS region was subsequently digested with DraI. All three fragments from the pFLUC-No SBS digestions were then ligated to the PCR product.

To generate pTRE-FLUC-SERPINE1 3' UTR or pTRE-FLUC-FLJ21870 3' UTR, pFLUC-SERPINE1 3' UTR was amplified using the primer pair 5'-GATACCGCGGATGGAAGACGCCAAAAACATAAAG-3' (sense) and 5'-GTCAGAATTCGCTTCTATTAGATTACATTTCATTCAC-3' (antisense), or pFLUC-FLJ21870 3' UTR was amplified using the primer pair 5'-GATACCG CGGATGGAAGACGCCAAAAACATAAAG-3' (sense) and 5'-GTCAGAAT TCGAGACAGAGTCTCCGTTGCCC-3' (antisense), where underlined nucleotides denote a SacII or EcoRI site, respectively in each primer pair. The resultant PCR product was digested with SacII and EcoRI and inserted into SacII- and EcoRI-cleaved pTRE vector (Clontech).

Cell culture, transient transfection and formaldehyde crosslinking. Human (HeLa or HaCaT) cells (2×10^6 per 60-mm dish or 7.5×10^7 per 150-mm dish) were grown in the medium DMEM (GIBCO) containing 10% FBS (GIBCO). Cells were transiently transfected with the specified plasmids by using Lipofectamine 2000 Transfection Reagent (Invitrogen) or with the specified siRNA by using Oligofectamine Transfection Reagent (Invitrogen) as previously described¹. The siRNAs used were *STAU1*-directed siRNA¹³, 1/2-sbsRNA1-directed siRNA (5'-CCUGUACCCUUCAGCUUACdTdT-3'), 1/2-sbsRNA1(A)-directed siRNA (5'-AUGACUUUGGGCAAAGUACdTdT-3'), *DICER1*-directed siRNA (Ambion), *AGO2*-directed siRNA (Ambion), *SERPINE1*-directed siRNA (Ambion), *FLJ21870*-directed siRNA (Ambion), *RAB11FIP1*-directed siRNA (Ambion), 1/2-sbsRNA2-directed siRNA (5'-GGUGCAAAGACAGCAUUCdTdT-3'), 1/2-sbsRNA3-directed siRNA (5'-UAGUAGUCAAGACCAAUUCUAdTdT-3'), 1/2-sbsRNA4-directed siRNA (5'-UGGCAUUCAGUUGAGUUUdTdT-3') and a nonspecific siRNA, *Silencer* Negative Control #1 siRNA (Ambion). Notably, all lncRNA-directed siRNAs used in this study target a sequence outside the Alu element. For all immunoprecipitations, cells were crosslinked using 1% formaldehyde for 10 min at 25°C and subsequently quenched with 0.25 M glycine for 5 min at room temperature before lysis¹. In experiments that blocked protein synthesis, cells were incubated with 300 $\mu\text{g ml}^{-1}$ cycloheximide (Sigma) 3 h before lysis.

For mRNA half-life measurements, Tet-Off HeLa cells (Clontech) were transfected with the specified siRNA in the presence of 2 $\mu\text{g ml}^{-1}$ doxycycline (Clontech). After 48 h, the medium was replaced to remove doxycycline, and cells were transfected with the indicated reporter and reference plasmids. After 4 h, an aliquot of cells was collected. Then, 2 $\mu\text{g ml}^{-1}$ doxycycline was added to the remaining cells to silence reporter gene transcription, and aliquots of cells were collected at time points thereafter.

Scrape-injury repair assays were essentially performed as previously published^{21,22}. Briefly, 2 days after transfection with siRNA, monolayer cultures of HaCaT cells at 90% confluence in 100-mm dishes were scratched in nine places using a P200 pipette tip (VWR) and uniform pressure to create denuded areas that were 0.9 mm wide. Cells were washed once with growth medium (DMEM supplemented with 10% FBS), which removes scratch-generated debris and generates smooth wound edges, and then cultured for an additional 16 h with monitoring.

Protein purification, immunoprecipitation and western blotting. HeLa cells were lysed, and protein was isolated using hypotonic buffer consisting of 10 mM Tris-Cl (pH 7.4), 150 mM NaCl, 2 mM EDTA, 0.5% Triton X-100, 2 mM benzamide,

1 mM phenylmethylsulphonyl fluoride and 1 tablet complete protease inhibitor cocktail in 50 ml (Roche). If the cells had been formaldehyde crosslinked, they were sonicated six times for 30 s to facilitate lysis. Immunoprecipitation was performed as previously described¹. In experiments that involved formaldehyde crosslinking, crosslinks were reversed by heating at 65 °C for 45 min after immunoprecipitation. Western blotting was performed as previously described¹. Antibodies consisted of anti-STAU1 (ref. 23), anti-calnexin (Calbiochem), anti-Flag (Sigma), anti-ILF3 (Santa Cruz Biotechnology), anti-FMR1 (Santa Cruz Biotechnology), anti-HA (Roche), anti-DICER1 (Santa Cruz Biotechnology), anti-AGO2 (Santa Cruz Biotechnology) and anti-BAG5 (Abcam) antibodies.

RNA purification, poly(A)⁺ RNA preparation and RT coupled to either semi-quantitative or quantitative real-time PCR. RNA was purified from total, nuclear or cytoplasmic HeLa-cell fractions or immunoprecipitated from total-cell lysates using TRIzol (Invitrogen) as previously described¹. Poly(A)⁺ RNA was extracted from total-cell RNA by using the Oligotex mRNA Mini Kit (Qiagen). Alternatively, RNA derived from different human tissues was obtained from Ambion. Semiquantitative RT-PCR and quantitative real-time RT-PCR were performed as previously described¹, using the designated primer pairs (Supplementary Table 4). In Supplementary Fig. 2c, RT was primed using oligo(dT)₁₈ rather than random hexamers. Semiquantitative RT-PCR analyses situated under the wedges in the leftmost lanes of figures involved twofold dilutions of RNA and show that the data fall within the linear range. RT-PCR values plotted as histograms include the standard deviation obtained in the specified number of independently performed experiments.

RNase protection assay and primer extension. For the RNase protection assay, the RPA III Ribonuclease Protection Assay Kit (Ambion) was used. Uniformly

labelled RNA probes (10⁷ c.p.m. µg⁻¹) were generated by transcribing linearized pcDNA3.1/Zeo(+)-Chr11_66193000–66191383 (which contains 1/2-sbsRNA1(S) and upstream and downstream flanking sequences) *in vitro* using [α -³²P]UTP (PerkinElmer) and the MAXIscript Kit (Ambion). Each probe (10⁵ c.p.m.) was incubated with poly(A)⁺ HeLa-cell RNA (10 µg) or yeast RNA (10 µg) in hybridization buffer (Ambion) at 42 °C for 12 h and subsequently cleaved using RNase A and RNase T1 (1/200; Ambion) at 37 °C for 30 min. Input probe (1/1000) and cleaved products were resolved in a 3.5% denaturing polyacrylamide gel and visualized using a Typhoon PhosphorImager (GE Healthcare).

Primer extension was performed using poly(A)⁺ HeLa-cell RNA (10 µg), SuperScript II reverse transcriptase (Invitrogen) and the 1/2-sbsRNA1-specific antisense primer 5'-GAGTTAAAAGAGGCTGCAGTG-3'. DNA sequencing was executed using the SILVER SEQUENCE DNA Sequencing System (Promega), the same antisense primer and pcDNA3.1/Zeo(+)-Chr11_66193000–66191383. Primer extension and sequencing products were resolved in an 8% denaturing polyacrylamide gel and visualized using a Typhoon PhosphorImager.

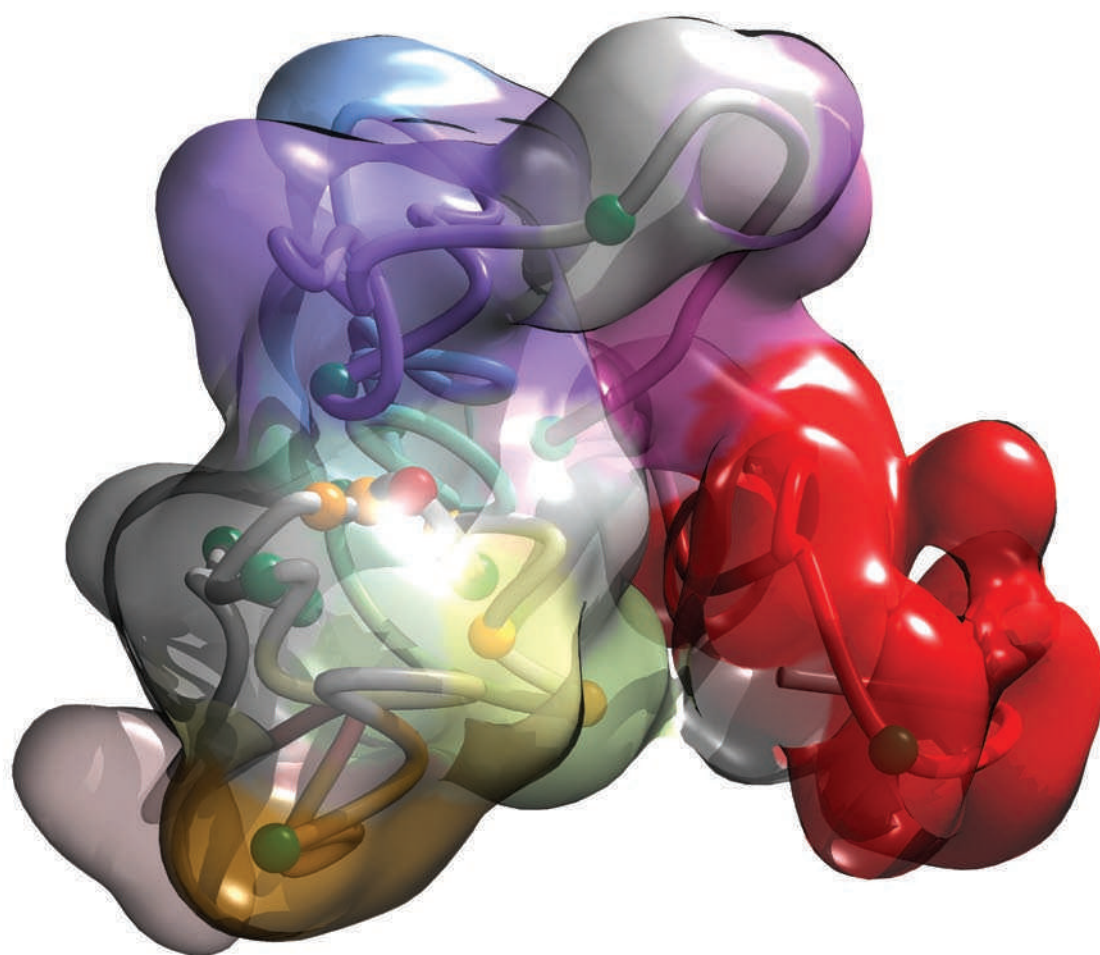
Fluorescence and phase contrast microscopy. Cells were visualized using an Eclipse TE2000-U inverted fluorescence microscope (Nikon), and a 480-nm excitation wavelength was used for phase contrast microscopy. Images were captured using TILLvisION software (TILL Photonics).

23. Marion, R. M., Fortes, P., Beloso, A., Dotti, C. & Ortin, J. A human sequence homologue of Staufin is an RNA-binding protein that is associated with polysomes and localizes to the rough endoplasmic reticulum. *Mol. Cell. Biol.* **19**, 2212–2219 (1999).

TECHNOLOGY FEATURE

GENOMES IN THREE DIMENSIONS

A DNA sequence isn't enough; to understand the workings of the genome, we must study chromosome structure.



Globular conformation of a 500-kilobase gene-rich domain on human chromosome 16.

BY MONYA BAKER

The next frontier of genomics is space: the three-dimensional structures of chromosomes coiled in the nucleus. Far from being the random result of packing 2 metres of DNA into a sphere perhaps 10 micrometres across, the structures vary across cell types and exert an as-yet-mysterious influence on gene expression. Efforts to decipher the effects of structure face many

difficulties, not least that researchers are still trying to find out how chromosomes shift as cells change, says Thomas Cremer, a geneticist at the Ludwig Maximilian University of Munich in Germany, who has studied the spatial organization of the genome since the 1970s. “The nucleus is still an uncharted landscape and it is embarrassing how little undoubtedly proven knowledge we have about its dynamic topography,” he says.

The basics have been known for decades:

DNA double helices coil around proteins called histones, forming ‘chromatin’ strands that in turn are bundled into chromosomes. But when it came to the twisting and turning of chromosomes themselves, “it wasn’t clear what role genome organization was playing or even if there was that much organization”, says Peter Fraser, a genome biologist at the Babraham Institute in Cambridge, UK. Long-range interactions seemed implausible. “People assumed that sequences 50 kilobases away couldn’t find

each other in the nucleus,” he says.

These days, scientists know that such interactions happen all the time. In 2002, Fraser’s laboratory was among the first to detect ‘long-range looping interactions’ that bring gene sequences into physical contact with far-off regulatory elements¹.

More-global changes also occur. For example, inactive chromatin is generally shunted to the nuclear periphery, but that arrangement is inverted in mouse retinal cells, allowing more light to reach photoreceptors². That the spatial organization of the genome is important is also demonstrated by the havoc that alterations can wreak. A cancer of the lymphatic system called Burkitt’s lymphoma occurs after a chunk of chromosome 8 ends up on chromosome 14 and vice versa. This happens because of the way that chromosomes arrange themselves in white blood cells³ — translocations occur more often between genes that physically come together during transcription⁴. Various types of cancer have been found to be connected with mutations in proteins that affect chromatin structure, and researchers have speculated that long-range interactions can be altered by disease-associated mutations in stretches of DNA that do not code for genes.

ANSWERS IN THE STRUCTURE

Researchers have long known that DNA sequences and histones are tagged with chemical modifications that turn genes on and off; the cataloguing of such ‘epigenetic’ modifications is well under way. It is now becoming clear that the three-dimensional organization of chromatin reflects a higher order of epigenetic regulation, says Yijun Ruan, a biologist at the Genome Institute of Singapore, who has developed techniques to find long-range interactions mediated by specific proteins⁵. Instead of assuming that gene activity is determined entirely by chemical attachments along a linear DNA sequence, researchers are looking for answers in the ways that chromatin folds, moves and communicates.

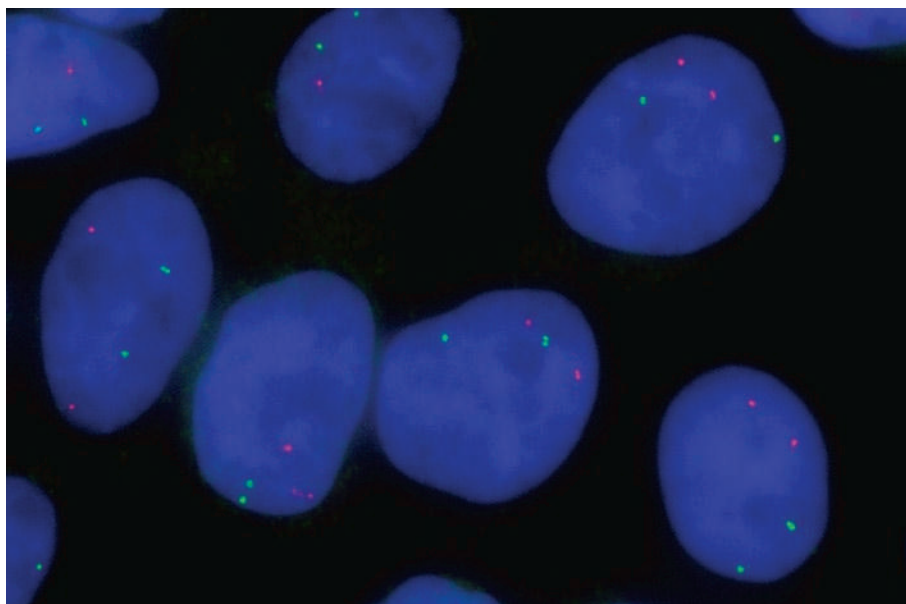
Discussions are beginning to include phrases such as ‘chromatin network’, ‘chromosome interactome’ and ‘spatial epigenetics’.

A suite of technological innovations is starting to reveal the significance of such concepts. New microscopes are letting researchers look more closely at more nuclei, for example, and experiments are allowing researchers to



“You start out with a difficult problem, and you convert it through a series of molecular steps to a simple problem.”

Job Dekker



Fluorescence *in situ* hybridization can illustrate the positions of genes within nuclei, such as the *MYC* (red) and *MMP1/3/12* (green) genes, shown here in breast tissue.

identify interacting sequences or to locate sequences within the nucleus. But challenges remain: chromosomal movements are dynamic and non-deterministic, so detecting what is where, and when, is difficult. Even more difficult is figuring out when and how genome architecture affects gene activity.

Until the beginning of this century, nearly all techniques that were used to study chromosome arrangements relied on microscopy. Researchers could label certain DNA sequences or DNA-associated molecules, and see where the labelled areas were inside the nucleus. But a strand of chromatin is only about 10 nanometres thick, and conventional fluorescence microscopy has a resolution at best of 200 nanometres. Thus, microscopy can reveal that two loci are close to each other, but not whether they come into contact. Moreover, if an interaction is fragile or short-lived, microscopy can miss it altogether.

When Job Dekker was a postdoctoral researcher studying the mechanics of cell division at Harvard University in Cambridge, Massachusetts, he wanted to map the DNA sequences that mediated interactions between chromosomes. One day, while commuting to his lab, he hit on the idea of capturing an interaction by chemically snagging two strands of chromatin that approached one another, then fusing the DNA from both into a single molecule. “You start out with a difficult problem — where are two loci in three dimensions — and you convert it through a series of molecular steps to a simple problem, just sequencing a piece of DNA,” says Dekker, now a genome biologist at the University of Massachusetts Medical School in Worcester.

Dekker’s idea became a technique, described in the literature in 2002, known as chromosome conformation capture (3C; ref. 6). It has since

spawned many variations (see ‘Investigating the architecture’), but the basic principles are the same. Protocols begin with ‘cross-linking’: dousing cells with formaldehyde to glue the DNA to its associated proteins, and those proteins to each other. Then the DNA is cut up with restriction enzymes or sheared by sonication, leaving behind ‘hairballs’ of tangled DNA and protein.

The next steps vary between protocols, but all combine free strands of DNA to create hybrid molecules: ligation products of DNA strands that had been close together on the same hairball. Researchers interested in genes that are associated with a particular transcription factor or other DNA-associated protein use specially designed antibodies to capture the relevant hairballs. In some techniques, chemically modified nucleotides are incorporated into hybrid molecules to ease purification, whereas in others, judicious application of PCR amplifies DNA sequences near loci of interest.

THE MEDIUM MATTERS

No matter which technique is used, researchers need to be careful when choosing their restriction enzymes. For example, those that cut at sites made up of 6-base-pair sequences produce large fragments that may not capture important interactions, whereas enzymes that recognize sequences of 4 base pairs may produce more and smaller fragments, perhaps generating so much background information that real interactions cannot be detected.

Researchers also need to keep in mind that most of the hybrid DNA molecules produced by this technique are the result of random interactions, particularly between loci that are just a few kilobases apart on the same chromosome; separating the signal from the background noise requires involved

MISTEL LAB

INVESTIGATING THE ARCHITECTURE

Technique	Detects	Protocol after sample is cross-linked and digested	Detection method
3C (chromosome conformation capture) ⁴	Interactions between two loci	Free DNA ligated. PCR is performed with one primer for each locus, and products are amplified	Quantitative PCR (qPCR)
4C (circular 3C; also known as 3C on a chip) ^{16,17}	Interactions between one locus and the rest of the genome	DNA molecules self-ligate into circles. Inverse PCR amplifies molecules containing loci of interest	Sequencing or microarrays
5C (3C carbon copy) ¹⁸	Interactions between multiple selected loci (for example, several within one chromosome)	Extension primers are designed for each interaction to be analysed. These allow selected DNA to ligate and be amplified by PCR	Sequencing or microarrays
ChIP-loop (chromatin immunoprecipitation loop) assay ¹⁹	Interactions between two loci bound by a particular protein	DNA bound to protein is purified and then ligated. PCR is performed with one primer for each locus, and products are amplified	qPCR
ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) ⁵	Genome-wide interactions for loci bound by a particular protein	DNA bound to protein is purified and then ligated. Ligation incorporates biotin-containing primers that allow ligation junctions to be purified	Deep sequencing
DamID (DNA adenine methyltransferase identification) ²⁰	Sequences that occur near nuclear landmarks or other proteins	Genetically modified cell lines are produced so that a DNA-tagging enzyme is fused to a protein of interest	Sequencing that detects methylated adenine
e4C (enhanced ChIP-4C) ⁷	A more-sensitive version of 4C that does not require inverse PCR	Primer extension with biotinylated primer that allows ligated fragments to be purified	Sequencing or microarrays
Hi-C (high-throughput 3C) ¹²	Genome-wide interactions at a resolution of about 1 megabase	Ligation step incorporates biotin-containing nucleotides that allow ligation junctions to be purified	Deep sequencing

bioinformatics and replicated experiments. “It used to be, even two years ago, that getting the data would be an endpoint of the project. Now it’s the start,” says Dekker.

On the plus side, preparing libraries of ligation products requires only very general reagents: formaldehyde, a variety of buffers and the enzymes that cut DNA and join it back together. Moreover, all the necessary reagents can be purchased from established companies: Life Technologies of Carlsbad, California; New England Biolabs of Ipswich, Massachusetts; QIAGEN of Hilden, Germany; Sigma-Aldrich of St Louis, Missouri; and Thermo Fisher Scientific of Waltham, Massachusetts. Researchers can also order specially synthesized primers for DNA amplification or ligation from a large range of (generally smaller) providers.

Different techniques generate different information. A million sequenced molecules (or ‘reads’) for Hi-C (high-throughput 3C) provides a low-resolution map of the whole human genome, whereas a million reads for 4C (circular 3C) produces a detailed interaction map for a gene of interest, and in ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) the same amount of data indicates which transcription-factor binding sites interact with which gene promoters.

This summer, Life Technologies plans to launch a kit that bundles together reagents for 3C experiments. The kit would allow

researchers to monitor and optimize digestion, use less of the sample for ligation and produce a library of ligation products in 1.5 days, says Shoulian Dong, a technology developer at Life Technologies. But perhaps the most important factor for throughput is the increasing availability of next-generation sequencers from companies such as Applied Biosystems of Carlsbad, California, and Illumina of San Diego, California, which can quickly sequence the hundreds of thousands of short hybrid DNA molecules produced in these experiments.

FROM SEQUENCES TO IDEAS

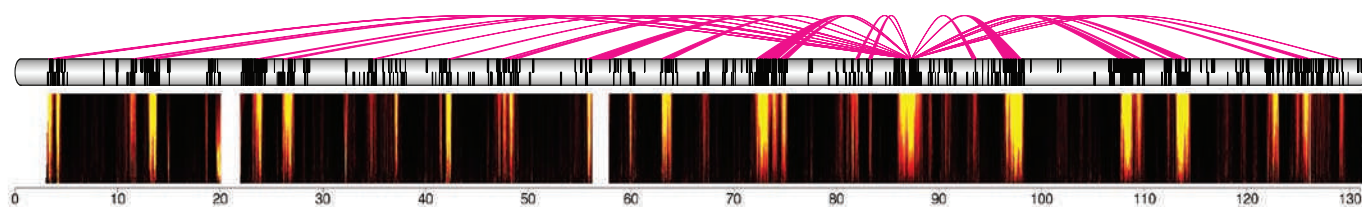
The ability to detect specific interacting loci is already revealing previously unknown biology. Last September, researchers led by Richard Young, a molecular biologist at the Massachusetts Institute of Technology in Cambridge, described evidence for a biological system that juxtaposes separate stretches of DNA. Together, these stretches control gene expression. The team found that a ‘mediator’ protein complex was often bound to enhancer sequences and core promoters of genes transcribed in embryonic stem cells⁷. Another protein, cohesin, which can connect two DNA segments, was bound along with mediator, and purified with it. Follow-up 3C studies on four genes showed increased interactions between promoter and enhancer sequences in stem cells, but not in

another type of cell in which the genes were inactive⁷.

For Wouter de Laat, a genome biologist at the Hubrecht Institute in Utrecht, the Netherlands, who showed how 3C can be used to match a gene with its regulatory elements⁸, the most exciting applications of chromosome capture technology are global: working out which sites interact with which genes in different tissues. “There are many more sites with regulatory potential than we have genes, and the only way to know which site is acting on which gene is to get three-dimensional,” he says. “That’s the next level of what we need in functional genomics.”

Current techniques are not powerful enough to match regulatory elements and genes across the genome, but de Laat and other labs are working on more far-reaching methods, which they hope to describe in the literature this year. It is useful to ask genome-wide questions because, otherwise, researchers tend to interpret their results only in the context of the gene they happen to be studying, says de Laat. But because every gene is part of a chromosome, those observations could have less to do with the gene under study than with its neighbours.

Adding to the challenge is another signal-to-noise problem: all the current techniques have to be carried out on between 10 million and 20 million cells at once, which means that the observed interactions represent an averaged



Regions on mouse chromosome 8 that interact with the *Rad23a* gene. The interactions were uncovered using conformation capture techniques.

reading. No one believes that all the interactions identified by sequencing technologies occur in any one cell, says Tom Misteli, who studies the cell biology of genomes at the US National Cancer Institute in Bethesda, Maryland. "Any interaction that happens will appear as a signal, but it doesn't tell you how often it happens in cells," he adds. "That makes the interpretation of the sequencing data a little bit complicated."

SEEING IS BELIEVING

To find out how often interactions occur, researchers have to count labelled cells under a microscope. For live-cell imaging, they can insert genes for fluorescent proteins that bind to desired DNA sites into the cell, but the technique is labour-intensive and tedious. A fixed-cell technique, fluorescence *in situ* hybridization (FISH), is more common. Nuclei are treated with formaldehyde, then denatured just enough to allow the entry of DNA probes that fluorescently label certain sequences.

In general, interactions identified by chromosome conformation studies are observed in only about one in ten cells under the microscope, says Misteli. That doesn't mean that the interaction isn't real; randomly selected loci are seen near each other even less often. Instead, such rates show just how dynamic and varied chromosome arrangements are, and how difficult they can be to study.

Last year, Fraser and his colleagues combined chromosome capture technology with microscopy to show that a single transcription factor, Klf1, helps to bring target genes from distant loci into a cluster in a common space⁹. Such studies of 'transcription interactomics' could reveal secrets of cell differentiation and stability, but mastering the necessary technologies is a formidable task. To separate relevant hybrid molecules from background signals, the researchers made significant tweaks to the 4C technique. And to show that multiple loci came together at the same time, lead author Stefan Schoenfelder looked at some 50,000 cells under a microscope: the equivalent, Fraser says, of spending half a year in a dark room.

That situation is familiar to Misteli, who in 2009 used FISH to show how genes reposition themselves in cancer¹⁰; such knowledge could aid diagnosis. Genes generally move from the periphery of the nucleus towards the centre when they become active, but individual genes move in unpredictable ways. No one has yet been able to look at gene positioning comprehensively, to discover how it might vary across different cell types, says Misteli. "It's all based on small sample numbers and people's favourite genes. So you want to look at more genes and that's simply not possible."

Technologies are improving, letting

researchers look at more cells; Fraser says that currently available microscopes with faster autofocus and more-agile robotic stages would now let Schoenfelder perform the same number of experiments in a month or less. Platforms are available: PerkinElmer in Waltham, Massachusetts, sells the Opera high-content screening system, which keeps the objective lens immersed in water. This allows it to work at the high resolutions required to determine where sequences are in the nucleus. The instrument automatically moves along wells on a plate to collect the necessary data, and its four different-coloured lasers can light up several probes in each cell.

The Opera instrument can examine loci in hundreds of cells a minute — considerably faster than stand-alone microscopes — and can make difficult techniques more accessible to non-experts, says Achim von Leoprechting, vice-president of imaging at PerkinElmer. "We're seeing FISH moving out of specialized

in a sugar solution, carefully freeze them, cut them into thin slices, then add DNA probes¹¹. The process is technically demanding but produces fewer artefacts and better resolution than standard FISH because the probes don't need to move through an entire nucleus. Pombo has used cryoFISH to show that chromosomes keep largely to their own 'territories' but intermingle extensively⁹.

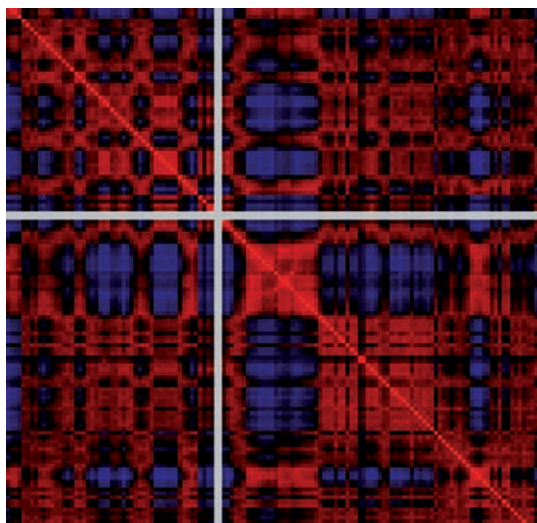
Electron microscopy has very high resolution, but the staining and imaging of cells can take days. In the past three years, researchers have turned to super-resolution optical microscopy, which uses techniques such as synchronized laser pulses to focus on structures as small as 15–20 nanometres — well below the 200-nanometre resolution limit of conventional optical microscopy — even in living cells. Companies selling these new microscopes include Applied Precision of Issaquah, Washington; Leica of Wetzlar, Germany; Nikon of Shinjuku, Japan; and Zeiss of Oberkochen, Germany, but the instruments have not yet reached most laboratories.

A THIRD WAY

Ultimately, all microscopy is a coarse detection technique, says Rolf Ohlsson, an epigeneticist at the Karolinska Institute in Stockholm. Standard fluorescence microscopy cannot distinguish between loci that are near each other and those that are in contact; even super-resolution microscopy cannot do so definitively. On the other hand, sequencing techniques cannot show which interactions occur together, says Ohlsson. "Somewhere between DNA FISH and chromosome conformation capture is the truth," he adds. But even accurate representations will not be enough: ascertaining that an interaction occurs is far easier than showing that it affects function. "Is what you see an interaction?" asks Ohlsson. "Or just a collision?"

Several groups are attempting to use conformation capture to build computational models that show the positions of chromosomes in different cell types and at different stages of the cell cycle. To construct these models, researchers do not actually measure distances between two loci; instead, they use algorithms to process captured DNA sequences. The programs produce 'proximity profiles' from sequencing data by measuring how frequently regions of the genome are observed to interact with one another, and comparing that with what would be predicted from chance.

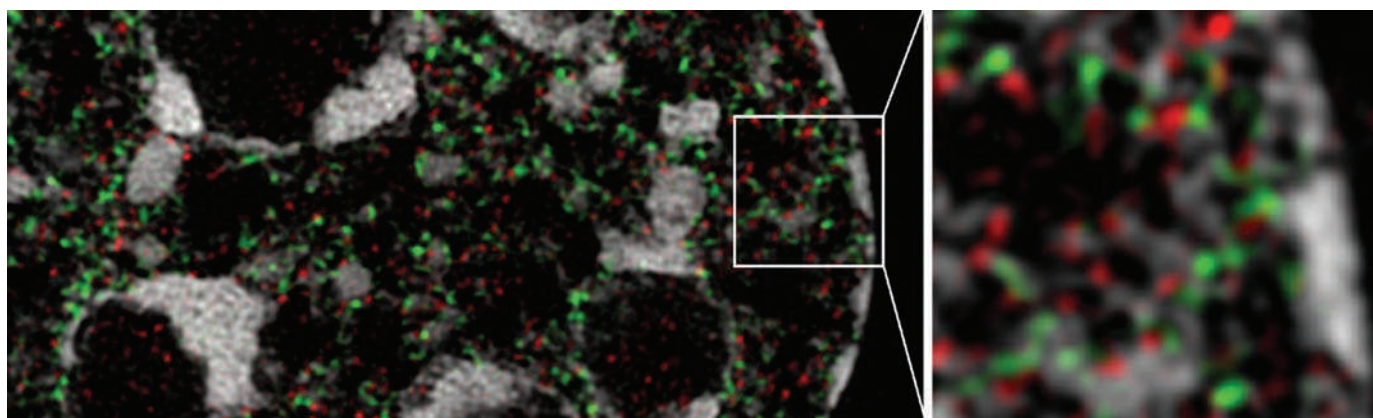
In 2009, Dekker and his colleagues constructed a model of human cells that breaks the 3-billion-base-pair genome into 3,000 pieces and maps long-range interactions¹². That resolution is too poor to show individual genes, let alone predict which binding sites might help to generate a particular conformation, but



Loci can interact more (red) or less (blue) than would be expected given their distance in the genome.

labs," he says, "so from an imaging standpoint we need to make sure they can use these platforms and get high-quality data without being trained as microscopists." Researchers who are already studying the position of genes in the nucleus are particularly keen to examine more cell types under different conditions, says Aaron Risinger, a specialist in high-content screening at PerkinElmer. "For individuals who were doing one-off experiments, the natural progression is to move to high-throughput," he says. In fact, Misteli is doing just that by incorporating the platform into a new US National Cancer Institute facility aimed at ultra-high-throughput cell biological imaging.

Lower-throughput techniques also have their advocates. Ana Pombo, a cell biologist at Imperial College London, has developed the cryoFISH technique: rather than fixing and denaturing intact cells, researchers embed cells



Super-resolution image of part of a mouse-cell nucleus, showing dense regions of chromatin separated by DNA-free channels. RNA production (red) and DNA replication (green) occur in a layer of decondensed chromatin on these domains. Strands of chromatin occasionally loop long distances between domains.

Y. MARKAKI/T. CREMER LAB.

creating a more detailed picture is difficult. Constructing the interaction map required some 30 million reads of fused DNA molecules; improving resolution by a factor of 10 (to 100-kilobase pieces) would require some 3 billion reads, because the number of reads required increases exponentially as the resolution improves linearly. Even so, Dekker and his colleagues' maps agreed with established ideas about chromosome territories, indicating that gene-rich areas lie close together.

WHOLE-GENOME MODELS

This year, researchers led by Dekker and Marc Marti-Renom, a bioinformatician at the Prince Felipe Research Centre in Valencia, Spain, published the results of 3C carbon copy (5C) performed on two different types of cell. They used the data to build a three-dimensional model of a 500-kilobase region of human chromosome 16 (ref. 13). This region contains a cluster of housekeeping genes active in most cell types, and another set of genes active in only some cells. Using interaction-frequency maps, the researchers generated chromatin models for both cell types. These predicted the existence of compact chromatin structures in which active genes were clustered. In the cells in which both sets of genes were active, the chromatin in the model folded into two 'globules'. In cells in which only the housekeeping genes were active, only one globule formed. FISH experiments confirmed the overall size and shape of this region of chromatin in individual cells.

It is possible to construct genome-wide models at higher resolution, by starting with smaller genomes. Last year, Ken-ichi Noma, who studies gene expression at the Wistar Institute in Philadelphia, Pennsylvania, and his colleagues took this approach, generating a very high-resolution genome-wide model of the fission yeast *Schizosaccharomyces pombe*, which has only three chromosomes, containing a total of about 14 million base pairs and 5,000 genes¹⁴. The researchers calculated how close different pieces of chromatin were to each other

by dividing the genome into sections of just 20,000 base pairs, and confirmed several results with microscopy. Earlier that year, a multilaboratory team had built a kilobase-resolution model of the genome of the budding yeast *Saccharomyces cerevisiae*, which has 16 chromosomes¹⁵.

The challenge starts with gathering reliable data: picking out real interactions from background reads. "The hardest step was going from sequence data to a set of interactions we could trust and interpret functionally. We had the data in hand for a year before the paper was published," says William Noble, a genome



"Any interaction will appear as a signal, but it doesn't tell you how often it happens. That makes it complicated."

Tom Misteli

biologist at the University of Washington, Seattle, who leads one of four labs that produced the budding yeast model. The structure provides a visual interpretation that the human brain can understand, says Noble, but that interpretation can be taken only so far. "The structure isn't introduced until the very end because we didn't want to base any of our conclusions on the structure itself," he says. Other researchers acknowledge that such models could be useful, but worry that they could be misleading. "When you say that two points are folded together, what's in between? We don't have the physical parameters to predict what's really happening there," says Ruan. The distance estimates from high-throughput data represent an "unrealistic average" that does not take into account that chromatin is in constant, often non-directed, motion, says Pombo. "You make protein structures when you crystallize a protein," she says. "Nuclei are not like that."

Model builders reply that in future,

representations will reflect the dynamic, semi-random movements of chromosomes, and that current versions can still be valuable, by showing overall tendencies. "By imaging you highlight the variability. By chromosome capture you highlight the commonalities," says Dekker.

But Cremer suggests that researchers should spend at least as much time with their microscopes as with their computers. Before people can really understand what high-throughput sequencing data tell us about higher-order chromosome arrangements, he says, the field needs many more descriptive studies. "One has to be very careful about making generalizations at this moment, and we need a lot more data." ■

Monya Baker is technology editor for Nature and Nature Methods.

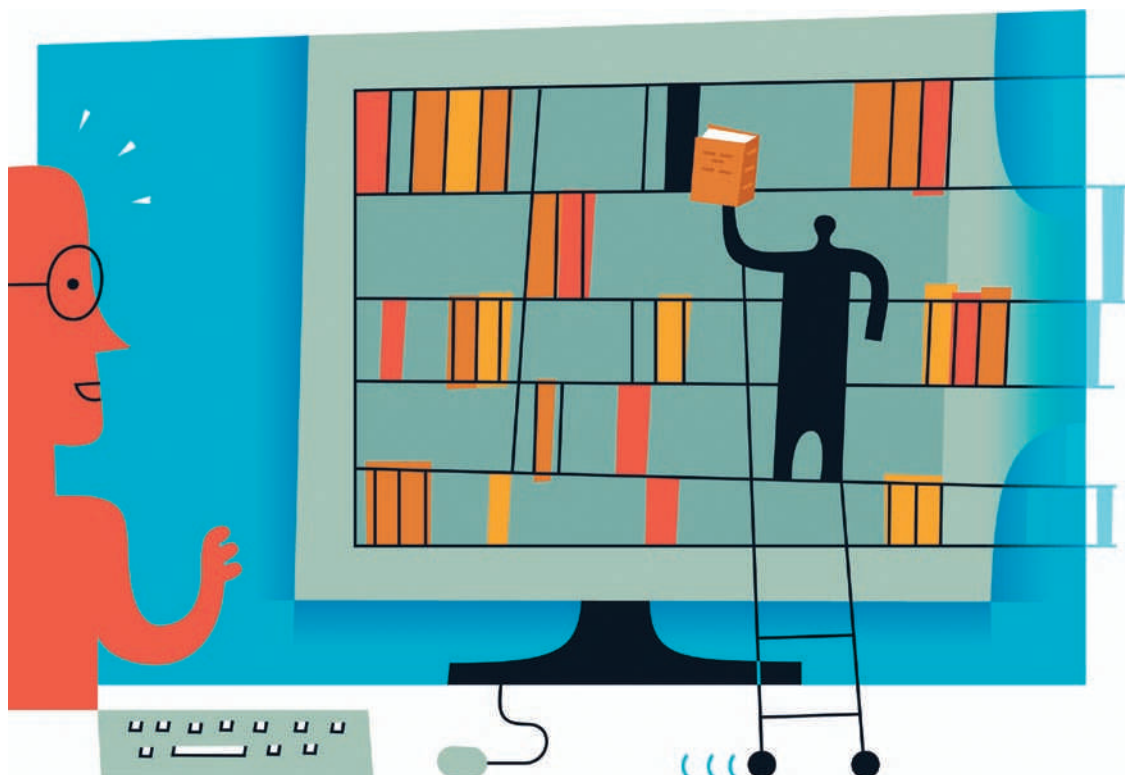
1. Carter, D., Chakalova, L., Osborne, C. S., Dai, Y.-F. & Fraser, P. *Nature Genet.* **32**, 623–626 (2002).
2. Solovei, I. et al. *Cell* **137**, 356–368 (2009).
3. Osbourne, C. S. et al. *PLoS Biol.* **5**, e192 (2007).
4. Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A. & Misteli, T. *Nature Genet.* **34**, 287–291 (2003).
5. Fullwood, M. J. et al. *Nature* **462**, 58–64 (2009).
6. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295**, 1306–1311 (2002).
7. Kagey, M. H. et al. *Nature* **467**, 430–435 (2010).
8. Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & de Laat, W. *Mol. Cell* **10**, 1453–1465 (2002).
9. Schoenfelder, S. et al. *Nature Genet.* **42**, 53–61 (2010).
10. Meaburn, K. J., Gudla, P. R., Khan, S., Lockett, S. J. & Misteli, T. *J. Cell Biol.* **187**, 801–812 (2009).
11. Branco, M. R. & Pombo, A. *PLoS Biol.* **4**, e138 (2006).
12. Lieberman-Aiden, E. et al. *Science* **326**, 289–293 (2009).
13. Bau, D. et al. *Nature Struct. Mol. Biol.* **18**, 107–114 (2011).
14. Tanizawa, H. et al. *Nucleic Acids Res.* **38**, 8164–8177 (2010).
15. Duan, Z. et al. *Nature* **465**, 363–367 (2010).
16. Simonis, M. et al. *Nature Genet.* **38**, 1348–1354 (2006).
17. Zhao, Z. et al. *Nature Genet.* **38**, 1341–1347 (2006).
18. Dostie, J. et al. *Genome Res.* **16**, 1299–1309 (2006).
19. Horiike, S.-I., Cai, S., Miyano, M., Cheng, J.-F. & Kohwi-Shigematsu, T. *Nature Genet.* **37**, 31–40 (2004).
20. Greil, F., Moorman, C. & van Steensel, B. *Methods Enzymol.* **410**, 342–359 (2006).

CAREERS

TURNING POINT Turkish academic post fuels US molecular biologist's career **p.297**

GRANTS US National Science Foundation requires grantee data-sharing plan **p.297**

NATUREJOBS For the latest career listings and advice www.naturejobs.com



BIOINFORMATICS

Curation generation

With biological databases growing in size and number, curators are needed to update and correct their contents. For those who prefer computers to pipettes, there are opportunities.

BY KATHARINE SANDERSON

Biologist and self-confessed bookworm Klemens Pichler thinks that he has found his ideal vocation. Pichler is a biocurator at the European Bioinformatics Institute (EBI) in Hinxton, UK, working on the Universal Protein Resource (UniProt) database. Some scientists would find it onerous to spend their days reading papers and sifting through and cross-referencing data. Pichler sees it as satisfying detective work, with a well-organized database as the result.

Biocurators are an unusual type of biologist. Their job is to make sure that the data such as

gene or protein sequences entered into large biological databases are standardized and annotated so that other biologists can understand them. "Once you have generated a sequence and identified a gene, there is an enormous amount of pre-existing data that you search that gene against. You need an expert to refine that information and make it usable," says Owen White, a bioinformatician at the University of Maryland School of Medicine in Baltimore. White developed the first genome-annotation software in 1995, and has been involved in several high-profile genome-sequencing projects.

At present, the number of biocurators is small — the International Society of Biocuration,

founded in late 2008, has just 300 members who work at some 100 organizations. But the number is likely to increase as sequencing becomes easier and biological data continue to roll in. By July 2008, more than 18 million articles had been indexed in the PubMed biomedical database, and nucleotide sequences from more than 260,000 organisms had been submitted to the GenBank database (see *Nature* **455**, 47–50; 2008). Started in 2008, the 1000 Genomes project has added to the data influx.

Pichler started work at UniProt after completing a fairly typical early academic career path: a degree in biology at the University of Vienna; postgraduate lab experience at Harvard ►

IMAGES.COM/CORBIS

► University in Cambridge, Massachusetts; and a PhD in virology at the University of Erlangen-Nürnberg in Germany, followed by a brief postdoc position there. It was during his postdoc that Pichler realized that he was on the wrong track. “I had grown tired of the frustrations of lab work,” he says. He read around and discovered biocuration; this was the change he had been looking for. “I’ve always been fond of computers but I never got round to integrating that into my career,” he says. Biocuration, Pichler found, was a way to make use of his training and move towards bioinformatics.

“It’s a wonderful career,” says Judy Blake, a bioinformatician at the Jackson Laboratory in Bar Harbor, Maine. Blake is a principal investigator on the Mouse Genome Informatics project, which employs 31 biocurators across multiple sites. She says that biocuration provides access to intellectual science without the stresses and responsibilities of finding funding and producing publishable results. Some researchers-turned-biocurators also relish the opportunity to be more of a generalist after academic careers that had a narrow scope.

PRACTICAL UNDERSTANDING

Although a PhD is not required, prospective biocurators need to be well trained in biology, with at least an undergraduate degree in a biological science and some related lab work. “Lab experience is important,” says Sandra Orchard, a senior scientific database curator at the EBI. “You can teach people curation but you can’t go back and teach them ten years at the bench.” Such experience helps biocurators to understand the data that they’re curating and how those data were generated.

Some universities offer specialist degree courses in biological information and the more software-design oriented bioinformatics, but none has a formal curation degree course specific to biological data. General data-curation programmes are available at the University of Illinois at Urbana-Champaign and the Digital Curation Centre in Edinburgh, UK, which offers short courses.

At UniProt, which employs almost 70 curators in Britain, Switzerland and the United States, Pichler spends half his time digging around to find out more about the protein sequences — the order of amino acids in a given protein — that are sent to the project from researchers around the world. He takes all the information he receives with each sequence

and compares it with existing entries in the database. He also does a thorough literature search. “You have to be a bit of a bookworm; you have to like reading and delving into matters and rummaging around and looking for clues,” says Pichler. He routinely scours the literature to find, for example, germane bits of information about the structure and function of a protein sequence. Next, he organizes and standardizes that information so others can interpret and understand it. “I concoct a new database entry, which then undergoes several rounds of quality control before it ends up being publicly available,” he says.

The other half of Pichler’s job is more technical, veering towards bioinformatics and software. He writes ‘rules’ so that computer programmes can annotate sequences with the structure and function of the genes or proteins. Researchers can then use these rules on their computers to predict protein function and structure from sequence data. Similar tasks are required for other databases, from those focused on gene-sequencing, such as Blake’s mouse-genome project, to efforts such as the Gene Ontology project, which aims to standardize gene representation across species.

The extent of curation depends on the database — the needs of a simple repository for information will differ from those of a comprehensive catalogue that combines information from direct submissions and published literature. Dealing directly with the scientists who produce the data — and can explain and modify the information on request — is easier than having to sift through the literature, says Orchard. “When working from a paper, you are dependent on it being well written in the first place and the data being complete and fully described. This is often not the case,” she says.

INTERNATIONAL COMMUNITY

Most large databases, and consequently curation jobs, are based in Europe and the United States, but that is changing, says Tadashi Imanishi, leader of the integrated-database and systems-biology team at the Biomedical Information Research Center in Tokyo, part of the National Institute of Advanced Industrial Science and Technology. The International Society of Biocuration has helped curators in Japan and other countries be part of the community. “By joining the society, they have the chance to communicate with curators in many other databases in the world,” says Imanishi, noting that Japan now has some 100 biocurators working on projects such as the DNA Database of Japan, which employs about 20 biocurators, and the H-Invitational, an international effort to catalogue all human genes.

At the moment, most jobs are at universities. But industry is beginning to offer biocuration services. For example, Ingenuity Systems in Redwood City, California, founded in 1998 by Stanford University graduate students, employs biocurators in its offices in Germany,

Switzerland, France, Britain and Japan. They look after the Ingenuity Knowledge Base, which the company claims is the world’s largest curated database of biological networks, documenting the relationships between proteins, genes, complexes, cells, tissue, drugs, disease and biological pathways.

Because of the skew towards academia, one of the biggest challenges to the growing field is its dependence on grant money. “Right now there is poor recognition for the value of curation,” says White. Funding agencies should

factor the cost of curation into grants, he says, although this can be difficult given tight budgets and the field’s relative infancy. “We’re in a very, very competitive market and have to work hard to justify curation to agencies,” he says. Yet, he adds, “this kind of librarianship is critical”. Sequencing may be increasingly cheap and sequenced genomes plentiful, but without curation the data mean little.

Although long-term funding can be

elusive, jobs can be lucrative. US biocurators in their first positions earn around \$65,000, says Blake — more than a postdoctoral researcher. In Britain, salaries start at around £31,000 (US\$48,000). And there is scope for advancement, says Orchard — a biocurator could end up running a database or training users. Curation also could be a doorway to computer programming and bioinformatics. Biocurators need not have any software-engineering expertise, but they do work closely with the people who write the programmes they use, and anyone interested in software design could move in that direction.

Blake says those considering a career in biocuration should know that it will move them away from the lab, which could pose a problem for those wishing to re-establish independent research, build a publication record or find grant funding. “None of these aspects is an integral part of the duties or outcomes of a biocurator position,” she says.

“There’s no doubt it’s a desk job,” Pichler concedes. But many don’t mind. They like the continued focus on science, as well as the occasional opportunity to attend conferences, give a talk or write an academic paper about their database, says Blake. “Curators,” she says, “do novel work that is required by everyone doing science.” ■



“Lab experience is important. You can teach curation but you can’t teach ten years at the bench.”

Sandra Orchard



“You have to like reading and delving into matters, rummaging around looking for clues.”

Klemens Pichler

Katharine Sanderson is a freelance writer based in Toulouse, France.

TURNING POINT

Cory Dunn

Cory Dunn, a US molecular biologist, celebrated his first year as an assistant professor at Koç University in Istanbul, Turkey, last September. Since arriving, he has helped to build a molecular biology department and has earned an installation grant from the European Molecular Biology Organization (EMBO).

How was college pivotal to your career?

I was studying biology at the University of Toledo in Ohio in preparation for studying medicine. I grew up in Ohio, had barely ever left the state, and planned to stay there. In my junior year, as part of an exchange programme, I was sent to the University of Salford, UK. Gaining a sense of another world of science overseas shook my vision of what was possible. I took upper-level courses there — biochemistry and molecular genetics. That led me to a science PhD at Johns Hopkins University in Baltimore, Maryland, and into research rather than becoming a doctor.

What is it about research that captivates you?

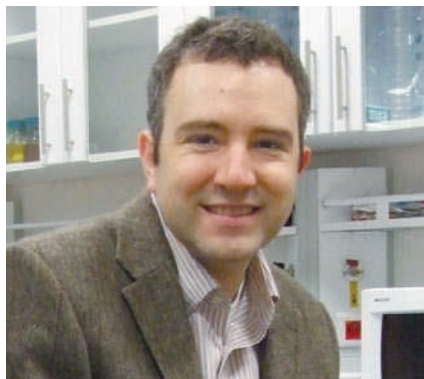
I want to learn new things. Making discoveries is like a game, and it's fun to design experiments. I'm always looking for the 'smoking gun' experiment to prove the point.

How did you come to move to Turkey?

After a postdoc at Columbia University in New York, I was eager to start my independent career. My wife, also a Johns Hopkins graduate, is from Turkey, and we decided to investigate research options there, because finding a position in the United States would require more postdoctoral work. US universities don't even want to see you without a National Institutes of Health new-investigator grant. The Turkish government has been pumping money into research and development. Koç University is a private university with a top-notch faculty. This was a big move, but not impulsive. They wanted someone to help found a department of molecular biology — a rare opportunity.

How did you convince them you were the right one to help found the department?

I told them about my interest, my family situation and my desire to meet in person. They said they wouldn't hire us as a couple, but my wife's record spoke for itself. She got a position. Last year, we hired our third faculty member — a Turkish citizen, coming from Harvard University in Cambridge, Massa-



chusetts. We're working to build the department, with two to three more hires in the next few months.

Are you going to focus on one topic, or hire for excellence?

I can see pluses and minuses to both approaches, but we decided that we want good people rather than a department based on, say, neuroscience or cancer. Simply hiring ambitious, smart people is best for morale. Interdisciplinary connections will come alive when you get the best people.

What has been the biggest challenge?

There is a lot of learning because this is a different culture, but it has been exciting to order equipment, establish policies and design classes. Everything is imported, so things can take longer when relying on import companies. But we've learned to plan ahead and work collaboratively. One thing that I have found challenging is learning Turkish. Everyone speaks English on campus so I haven't been pushed, but I am learning slowly.

How will the EMBO grant help your research?

External funding is good for our university as we try to promote opportunities. The research profile of Turkey is not well recognized, but there are good people doing molecular biology, chemistry and physics. So being able to talk about what's going on here, and encourage people to check it out, will help us to make connections in Europe.

Do you think you'll be in Turkey for long?

You can't predict the future, but we're happy here. We're doing everything to make this the best molecular biology department in Turkey, with a shining reputation in Europe. ■

INTERVIEW BY VIRGINIA GEWIN

EMPLOYMENT

US scientists keep jobs

Unemployment rates for US biological and physical scientists remains low compared with rates for the general population, according to 2010 data from the US Bureau of Labor Statistics (BLS). In the Current Population Survey, a poll of 60,000 households conducted by the US Census Bureau, geoscientists and environmental scientists reported 2.2% unemployment; chemists and materials scientists 3.1%; medical scientists 4.1%; and biologists 4%. Rates for each occupation in 2009 were 4.6%, 4.5%, 4.2% and 3.5%, respectively. The average rate of joblessness for the general population in 2010 was 9.6%. Richard Freeman, an economist at Harvard University in Cambridge, Massachusetts, says scientists with doctorates are much more likely to be employed than are those with only bachelors' or masters' degrees.

US NATIONAL SCIENCE FOUNDATION

Data policy takes effect

The US National Science Foundation (NSF) has implemented a 'data management plan' for its grant applicants, in effect from 18 January. Applicants are asked to specify, in no more than two pages, how data generated through their grants will be accessed, archived and shared — this includes revealing the types of data and other materials that will be produced, creating policies for data distribution and plans for archiving, and making provisions for accessing and sharing the data. These could, for example, address confidentiality and intellectual-property concerns. Announced last May, the guidelines let individual NSF divisions tailor the policy to their discipline's needs.

AGRICULTURAL SCIENCE

French recruitment

The French National Institute for Agricultural Research (INRA) in Paris is recruiting 50 junior scientists from around the world to develop healthy and sustainable food systems, mitigate greenhouse-gas emissions and adapt agriculture and forestry to climate change. The hirings are part of a scheme to recruit and retain junior and senior researchers. Recruitment ends on 24 February; applicants should have a doctorate and preferably a postdoc, says Thierry Boujard, INRA director of human resources. Those hired will become civil servants with starting salaries of US\$36,000–\$44,600. Tenure is possible after a year.

ESP

Breakfast with the enemy.

BY JULIAN TANG

The interrogation room was a disgrace. Its once shiny titanium walls and floor were stained with patches of unidentifiable dried goo. Commander Maurice Gilet sat to one side, waiting. A loud clattering and the thud of heavy equipment announced the arrival of the prisoner outside the room's entrance.

The door opened and Maurice's old friend, head prison guard Bernard Marchand, entered carrying an e-clipboard. "Prisoner AX-5777, as requested, Sir. Just transferred from holding at the Virgin leisure colony on Maldives-592."

"Thanks, Bernard — you can drop the 'Sir,'" he grinned, weakly. It had been a long week. "This is our prime suspect in the Virgin cruiser explosion?"

"Yes, but you've not interviewed one of these before, have you? They're totally aquatic, so the translation unit has given it a human voice — you'll approve, I think." He gave a wink and backed out of the room.

Maurice walked round the large, cylindrical water tank that held his captive. He stared at the contents curiously, and not without some amusement. The prisoner looked like a giant sea anemone. He glanced at the translation unit hovering beside the glass to check that it was functioning correctly, and sat back down.

"Do you like what you see, Commander Gilet?"

They had given it the voice of Audrey Hepburn, his favourite actress of all time. Yet, rather disconcertingly, he had also heard the voice in his head. Then another strange thing began to happen. His image of the sea anemone in the tank wavered, blurred, then disappeared, to be replaced by an image of Audrey Hepburn herself, sitting elegantly on a stool in her famous black *Breakfast at Tiffany's* dress, complete with diamond necklace, long black gloves and cigarette holder. He could even smell the smoke from the cigarette, as well as her perfume.

"Or would you prefer this?" she purred.

Maurice leaned back in his chair and rubbed his eyes. What the hell was this? No one had warned him that these creatures were telepaths. He shook his head to clear it.

"May I remind you that this is an interrogation and that this conversation is being recorded?" He struggled to make his voice sound authoritative. "You were found, drifting, among the crash debris, in a specially

adapted survival capsule. Nothing else survived the explosion. We want to know what happened and if, and how, you were responsible. Many people died in this explosion, so if you refuse to cooperate fully, things may become ... unpleasant."

Her seductive demeanour changed abruptly. "Commander, you are not in any position to caution me," she began in a tone of suppressed rage. "Your species has invaded our ecosphere purely for pleasure. The fact that the only intelligent life forms there are aquatic still does not permit the effluent from your 'terrestrial pleasure farms' to pollute our waters. Besides, you should be thanking me."

Maurice sighed and took the bait. "And why should we be doing that? Is this some sort of confession?"

She looked at him with an amused smile. "Commander," she began casually, "as you've seen, with my telepathic ability I can quite easily make any member of the crew load an explosive device on board, effectively by-passing any security. Yes, this is a confession, but perhaps you might like to ask yourself — why?"

Despite his growing irritation, Maurice grudgingly waved her on.

"As you know, there are no terrestrial life forms on our planet. This is because we have an ancient parasite that infected and slowly mutated these life forms until they became sterile — ultimately making them all extinct. Eventually, it adapted itself to water, where we have been monitoring its evolution very carefully over many years. Quite frankly, we fear this organism."

"Unfortunately, your largely aqueous human body is an ideal host for this parasite. Since you started colonizing our planet, the creature has been reverting to a more terrestrially adapted genotype and phenotype. Obviously, the organism does not show up on your

➔ **NATURE.COM**
Follow Futures on
Facebook at:
go.nature.com/mtoodm

routine environmental scans, but we detected the first infected humans aboard this vessel. Infected terrestrial hosts



usually exhale the organism continuously, making for highly efficient airborne transmission. In time, it would have infected your whole population. We will not allow this parasite to spread to other worlds."

Maurice was stunned. "So, what are you saying? That you destroyed an entire cruiser as some sort of infection control measure? If you had discussed this with us when we first starting building these leisure colonies, we could have worked together to develop a cure or vaccine then!"

Her image shimmered slightly and she transformed into her all-black, *Funny Face* leotard. Pouting, she continued. "Well, let's just say that your commercial developers were not particularly amenable to such an open dialogue." She paused, thoughtfully. "If we allow you to work with us on this, what about your environmental pollution? Although, admittedly the ammonia component

may be a useful contribution to our ecosystem, you will have to filter out the rest. Can this be done?"

Maurice briefly considered the request. "Yes, I will talk to the management. Given the alternative that you have very effectively demonstrated, I think they will listen."

She wasn't finished. Stretching languorously like a lean, black cat she added: "And you will, of course, limit the numbers of visitors?"

Maurice started laughing. They were not so different from humans, after all. "I'm sure that can be negotiated — given your particular talents."

"Well then, Commander, I think this is the beginning of a beautiful friendship."

And when he looked again, all he saw was a huge sea anemone with its tentacles gently waving in a large cylindrical tank. The translation unit was now sitting beside it, in sleep mode. ■

Julian Tang is a clinical/academic virologist, who still has many story ideas. He would like to dedicate this story to Audrey Hepburn, whose iconic elegance, although often imitated, has yet to be surpassed.

JACEY